Enabling pervasive applications by understanding individual and community behaviors

Lin Sun

THESE DE DOCTORAT CONJOINT TELECOM SUDPARIS et L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Lin SUN

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Nouvelles applications pervasives par la modélisation des comportements individuels et communautaires

Soutenue le 12 Décembre 2012

devant le jury composé de : (préciser la qualité de chaque membre)

François Brémond	Rapporteur	Professeur, INRIA – Sophia Antipolis – France
Jianhua Ma	Rapporteur	Professeur, Hosei University – Japan
Jean-Michel Dumas	Examinateur	Professeur, Université de Limoges – Limoges – France
Thierry Artières	Examinateur	Professeur, UPMC – Paris – France
Mounir Mokhtari	Directeur de thèse	Professeur, Institut Mines-Télécom – Evry – France
Daqing Zhang	Codirecteur de thèse	Professeur, Institut Mines-Télécom – Evry – France

Thèse n° 2012TELE0053

Doctor of Philosophy (PhD) Thesis
Université Pierre & Marie Curie -TELECOM SudParis

Specialization

Computer Science

presented by

Lin SUN

Submitted for the partial requirement of

Doctor of Philosophy
from
Université Pierre & Marie Curie (UPMC) - TELECOM SudParis

Enabling Pervasive Applications by Understanding Individual and Community Behaviors

Dec. 12, 2012

Commitee:

François Brémond	Reviewer	Professor, INRIA - Sophia Antipolis - France
Jianhua Ma	Reviewer	Professor, Hosei University - Japan
Jean-Michel Dumas	Examiner	Professor, University of Limoges - France
Thierry Artières	Examiner	Professor, UPMC – Paris - France
Mounir Mokhtari	Thesis Director	Professor, Institut Mines-Télécom – Evry - France
Daqing Zhang	Advisor	Professor, Institut Mines-Télécom – Evry - France

Dedication

I firstly express my deepest thanks to my supervisors, Professor Daqing Zhang and Professor Mounir Mokhtari, who gave me the opportunity to do this thesis, and have been providing the help, suggestions and encouragement in all the time during the thesis. I will also thank Professor Shijian Li, who provided the taxi GPS dataset and all the testers in my experiment, thank you for the help.

My dearest family is the greatest support during my research in these years. Although my parents are farmers who never had the opportunity to go of the small village to see how broad the whole world is, their strong believe in me was the strength that drove me to where I am now. And also I would like to thank my dearest sister, who always took care of the family so that I could do my research without worries.

I would like to thank my colleagues who ever worked with me in my lab, including Dr. Pablo Castro, Chao Chen, Dr. Bin Li, Professor Bin Guo, Dr. Mossaab Hariz, Nan Li, Yang Yuan, Zhibo Lin, Professor Shijian Li, Dingqi Yang, Zhu Wang, Haoyi Xiong, Dr. Daqiang Zhang, Dr. Zhiyong Yu, Chanaphan Fon Prasomwong and others who worked temporarily with me. Thank you for the help you gave to me.

I would like to express my gratitude to Madame Françoise Abad, Professor Djamal Zeghlache, Madame Xayplathi Lyfoung and Dr. Abdallah Mhamed, for their kindness and help in these years.

In the end, I want to express my thanks to all the friends I made in Telecom SudParis for their kindness and help in these four years.

Abstract

With the rapid advancement of sensing, computing, networking and storage, recent years have been witnessing a phenomenal growth of sensor-embedded mobile phones and prevailing use of GPS devices in vehicles. The digital footprints collected from such systems provide novel ways to perceive an individual's behaviors. Furthermore, large collections of digital footprints from communities bring novel understandings of human behaviors from the community perspective (community behaviors), such as investigating their characteristics and learning the hidden human intelligence. The perception of human behaviors from the sensing digital footprints enables novel applications for the sensing systems.

However, it's not easy to recognize individual and community behaviors from the digital footprints. Firstly, there are various problems in the sensing data that need to be preprocessed. Secondly, there are big gaps to be bridged between the raw sensing data and the recognition of human behaviors. Thirdly, with the observation of behaviors in a community, learning the hidden human intelligence is a complicated problem. Lastly, for a smart sensing system that monitors the behaviors of a large number of individuals, it's often hard to be scalable and support real time responses.

Bases on the digital footprints collected with accelerometer-embedded mobile phones and GPS equipped taxis, in this dissertation we present our work in recognizing individual behaviors, capturing community behaviors and demonstrating the novel services enabled. In recognizing individual behaviors, we present the recognition of an individual's physical activities with the accelerometer readings collected from mobile phones placed in the pockets around the pelvic region. Particularly, to overcome the variance of location and orientation problem, we introduce the orientation invariant sensing magnitude to the sensor readings and improve the recognition accuracy about 8 percent. Besides, we also reduce the computing cost with a feature reduction method. With the GPS footprints of a taxi, we summarize the individual anomalous passenger delivery behaviors and improve the recognition efficiency of the existing method *iBOAT* by introducing an inverted index mechanism. Besides, based on the observations in real life, we propose a method to detect the work-shifting events of an individual taxi.

With real-life large-scale GPS traces of thousands of taxis, we investigate the anomalous passenger delivery behaviors and work shifting behaviors from the community perspective and exploit taxi serving strategies. We find that most anomaly behaviors are intentional detours and high detour inclination won't make taxis the top players. And the spatial-temporal distribution of work shifting events in the taxi community reveals their influences. While exploiting taxi serving strategies, we propose a novel method to find the initial intentions in passenger finding. With the strategies modeled for thousands of drivers, we isolate the good and bad ones by learning from good drivers, measuring the correlation between each strategy and performance, and classifying different driver groups.

Furthermore, we present a smart taxi system as an example to demonstrate the novel applications that are enabled by the perceived individual and community behaviors.

Keywords

large scale digital footprints, taxi digital traces, activity recognition, individual behaviors, community behaviors, human behavior analysis.

Résumé

Avec les progrès rapides de détection, de l'informatique, des réseaux et du stockage, ces dernières années ont été témoins d'une croissance phénoménale de téléphones mobiles avec capteurs intégrés et prévaut l'utilisation des appareils GPS dans les véhicules. Les empreintes digitales recueillies par ces systèmes offrent de nouvelles façons de percevoir les comportements d'un individu. En outre, de grandes collections d'empreintes numériques des communautés apportent de nouvelles compréhensions des comportements humains, telles que les enquêtes sur leurs caractéristiques et l'apprentissage de l'intelligence humaine cachée. La perception des comportements humains à partir des empreintes digitales de détection permet de construire des nouvelles applications sur les systèmes de détection.

Cependant, il n'est pas facile de reconnaître les comportements individuels et collectifs depuis des empreintes digitales. Tout d'abord, des divers problèmes dans les données de détection doivent être prétraité. Deuxièmement, il y a de grandes lacunes à combler entre les données brutes de détection et la reconnaissance des comportements humains. Troisièmement, l'observation des comportements d'une communauté et l'apprentissage de l'intelligence humaine cachée restent des problèmes complexe. Enfin, pour un système intelligent de détection qui surveille les comportements d'un grand nombre de personnes, il est souvent difficile d'être évolutive et de soutenir les réponses en temps réel.

D'après les empreintes digitales recueillies avec l'accéléromètre embarqué dans les téléphones mobiles et les taxis équipés avec GPS, nous présentons ici notre travail sur la reconnaissance des comportements individuels, la capture des comportements communautaires et la démonstration des nouveaux services activés. En reconnaissant les comportements individuels, nous présentons la reconnaissance des activités physiques d'une personne avec les lectures de l'accéléromètre recueillies à partir des téléphones mobiles mis dans les poches autour de la zone pelvienne. En particulier, pour surmonter le problème de la variation de l'emplacement et de l'orientation, nous introduisons la grandeur de détection qui est invariante avec l'orientation dans les lectures de capteurs et améliorons la précision de la reconnaissance d'environ 8%. Par ailleurs, nous réduisons également le coût de calcul avec une méthode de réduction de caractéristique. Avec les empreintes GPS d'un taxi, nous résumons les comportements anormaux du transport des passagers pour un individu et améliorons l'efficacité de la reconnaissance de la méthode existante IBOA' en introduisant un mécanisme de l'index inversé. En outre, sur la base des observations dans la vie réelle, nous proposons une méthode pour détecter les événements de changement de service d'un taxi individuel.

Avec des traces GPS à grande échelle et à l'aide des milliers de taxis, nous étudions les comportements anormaux pour le transport des passagers et les comportements de changement de travail et exploitons les stratégies de service de taxi. Nous constatons que la plupart des comportements anormaux sont des détours intentionnels et l'inclinaison détour élevé ne fera pas des taxis les meilleurs joueurs. Et la distribution spatio-temporelle des événements de changement de travail dans la communauté de taxi montre leurs influences. Durant l'ex-

ploitation sur les stratégies de service de taxi, nous proposons une méthode pour trouver les intentions initiales de la recherche des passagers. Avec les stratégies modélisées pour des milliers de chauffeurs, nous isolons les bonnes et mauvaises stratégies. Par l'apprentissage auprès des bons chauffeurs, par la mesure de la corrélation entre chaque stratégie et la performance, et la classification des différents groupes de chauffeurs.

En outre, nous présentons un système intelligent de taxi comme une étude exemplaire des nouvelles applications qui s'appuie sur les comportements perçus individuelles et communautaires.

Mots-clés

une grande collection d'empreintes numériques, de traces numériques des taxis, la reconnaissance d'activité, les comportement individuels, les comportements communautaires, l'analyse du comportement humain

Table of contents

Chapter 1

Introduction

Contents

1.1 Overview

With the rapid advancements of sensing, computing, networking, and storage, recent years have been witnessing a phenomenal growth of sensor-embedded mobile phones, prevailing use of Global Positioning Systems (GPS) in private and public transportation vehicles, and wide adoption of sensor networks in facilities and outdoor environments. According to IDC's figures for worldwide smartphone unit sales and market share, in the second quarter of 2012, there were 104.8 million Android and 26 million iOS smart phones sold world wide. And in 2008, there were 15.1 million GPS units sold in U.S. only. These systems provide rich services that bring benefits and conveniences to people. The GPS module in mobile phone and GPS navigation devices can provide the location information, search points of interests nearby (such as restaurants, hotels, shopping locations and petrol stations), and find ways to people's destinations. Besides, with the easy distribution and access of applications in app shops (e.g., Apple Store or Google Play), smart mobile phones become powerful computing platforms providing services far beyond the phone calls and

messaging. Supported by the embedded sensors like accelerometer, video camera, audio, GPS, proximity and even NFC, the powerful processors and rich memories, and network connections like 3G and WIFI, they become mobile computing platforms which possess the computing power of the traditional PCs, the mobility and the sensing capability. They are used for social networking, such as taking and sharing information like photos and texts in twitter and Facebook, surfing internet with mobile web browsers, enjoying multimedia contents like the music, movie and on-line videos, mobile office working such as dealing with emails and documents.

All these systems have led to an unprecedented accumulation of digital footprints — the digital traces people left during the interactions with the cyber-physical world [124]. Particularly in this thesis, we are interested in the digital footprints left by the sensing devices. Take the mobile phones for example, the embedded sensors can sense and record the surroundings all the time. The accelerometers record the acceleration signals of carriers' movements. And the GPS sensors collect the locations where they are and thus the traces they follow. The bluetooth sensors detect the nearby bluetooth devices which implicitly inform who are close to them. As for vehicles, the GPS navigators could potentially record the routes they follow and the speeds which in most cases are the reflections of the traffics. Modern smart taxi dispatch systems gather the real time GPS locations and passenger status of a large taxi fleet for efficient dispatch and thus can obtain the digital traces of a drivers community.

The collection of such digital footprints provides valuable opportunities for researchers to perceive human behaviors. Individually speaking, the collection of digital footprints of an individual records rich information about one's movements, behaviors and locations, and potentially be able to infer high level knowledge such as one's habits, preferences and daily routines. For example, experiments in lab environments already proved that, with accelerometer-embedded sensor boards attached on human body, we can recognize an individual's behaviors such as walking, running, driving and so on [8, 41, 51, 52, 69, 71, 72, 90, 116]. The GPS traces recorded by an individual's mobile phones can tell where one has been, how long one stayed and even what commune tools one took by data mining methods [25] and thus provide a comprehensive picture about ones' daily lives. Not only mobile phones, the GPS traces of vehicles such as taxis reveal where one taxi driver has been and thus which area one serves the most for one's business. And enhanced by the passenger status, we can feature an individual's passenger finding and delivery behaviors, and detect whether one behaves anomalously or not [17, 125].

Furthermore, the aggregation of digital footprints of large number of individuals provides novel ways to explore human behaviors from a community perspective. Unlike the individual behaviors which feature only one's private life, the large aggregation of human

behaviors in a community provides novel perspectives to understand them in a collective manner. For instance, with the GPS traces of the tourists in a city, we can easily see the popular tour route, which kinds of hotels and restaurants they stay, which kinds of commune tools they usually take and etc. Thus we can obtain a general picture of the tourist business.

Besides, we can also infer high level intelligence from the human behaviors in a community. As practically human behaviors exert impacts on certain result in an explicit or implicit way, the large aggregation of them provide possibilities to comprehend the internal relations and thus generate novel understandings about the human behaviors. For example, with the aforementioned GPS traces of the tourists, we can see their traveling itinerary (the tour route, the time stayed in each place, the commune tools) and the resulted experiences (Are their traveling route covers most of the tourist sites? How much time do they stay in the commune, Are their behaviors complies with what the majority people do?). And thus we are able to infer the optimized traveling route, organizing the optimized tour route and planning suitable time to stay in every tourist site. Such intelligent information is hard to obtain if merely looking at an individual's trace. In this paper, we denote the human behaviors which are explored from the community perspective instead of individually as community behaviors. This thesis presents our research in finding the common characteristics of the community behaviors and extracting the hidden intelligence from a large collection of sensing digital footprints from a community.

Meanwhile, the perception of the individual and community behaviors allows sensing systems to enable various novel pervasive applications. For example, the explore of the individual and community behaviors in the aggregated GPS traces of the tourists provides personal management for the individual traveller, helping one to plan the traveling route, tag the recorded photos and texts with the locations and recall the memory in the future. Moreover, it's also useful for the traveling agency to efficiently look for their guests, and the planning of the optimized tour bus route. To list a few other examples, with the activity of daily living recognized by mobile phones, we can measure how long a person is sitting (stationary) down in working environment. If one sits down for too long time(say, longer than 50 minutes), to keep health, he can be alerted with a mobile phone application which recommends to take a rest or walk around. To increase the awareness of the sedentary lifestyle and motivate people to participate in more exercise, it's possible use the exercise amount as input in entertainment games. For example, a mobile phone game named UbiFit garden [1] uses the daily exercise of people to control the growth of flowers in a virtual garden and effectively improves the exercise situations of the users [22].

There are already many studies about the individual and community behavior enabled

1. http://dub.washington.edu/projects/ubifit

applications based on the GPS traces collected from a taxi fleet. We can easily detect the pick-up and drop-off events and use the extracted the pick-up hotspots to guide the passenger finding process for taxi drivers [15, 56, 63, 86, 102] and the drop-off hotspots to direct the taxi finding for passengers [84]. Besides, we can also use the passenger delivery trips to estimate the traveling time among different areas in a city and monitor whether taxi drivers follow normal routes in passenger deliveries [17, 125]. Considering taxis as mobile sensing nodes that perceive the activities of city traffics, we can obtain the real time traffics [92,121], detecting the speeds and traffic accidents, and provide them to navigators or directly to the public for people's awareness.

However, it's not easy to recognize individual and community behaviors from the digital footprints. Firstly, there are various problems in the sensing data that need to be preprocessed. Normally the sensing data has various kinds of noises or deficiencies which require data filtering work. And in case of low sampling rate, techniques like data augmentation are needed to compensate the missing samplings. For example, when recovering the traveling route, the samples in the GPS trace may only cover a few nonadjacent road segments and cause difficulties to estimate the missing ones between them. Secondly, there are big gaps need to bridged between the low level raw sensing data and the high level human behaviors. For instance, when studying human behaviors by the acceleration of some body location measured by the placed mobile phones or sensor boards, it's hard to tell the activities directly based on the sensor readings. We need to uncover and extract the easy-to-deal-with features which are hidden intrinsically in the sensing data and are different from other activities and then use proper methods to differentiate them. Thirdly, with the observation of community behaviors, how to learn the hidden human intelligence is a great challenge. For example, with the passenger finding behaviors of all taxi drivers, how to tell generally which kinds of behaviors are good and which are bad is not an easy task to deal with. Lastly, from the system point of view, to handle the request of a large number of individuals, the systems are often required to be scalable and support real time responses. For instance, in the anomaly passenger delivery detection service presented in this thesis, the system has to deal with thousands of occupied taxis, tracking their traces and judging whether the drivers behave normally or not in real time.

In this dissertation we study the individual and community behaviors from the digital footprints collected with accelerometer-embedded mobile phones and GPS supported taxis. In studying the individual behaviors, we present the process of recognizing people's physical activities with the digital footprints collected from their mobile phones when placed in the pockets around the pelvic region. Unlike the previous experiment settings of attaching sensor boards in certain positions or orientations, the mobile phones are in the natural orientations within several possible pocket locations. So our challenge is how to accurately

recognize people's activities when the mobile phone has varying position and orientations. We conduct experiment to collect the accelerometer readings from 7 volunteers while they perform the physical activities and transform the data into feature vectors and then recognize the physical activities with data mining methods. Besides we propose a novel way to improve the recognition accuracy and reduce the computing cost so that they can be targeted at mobile phone platforms.

With the GPS footprints of taxis, we reveal the anomalous passenger delivery behaviors of an individual driver and our work in improving the recognition efficiency, which is critical for the real time responses of the targeted anomaly passenger delivery monitoring applications. How to recognize the anomalous passenger delivery trajectories is firstly studied in our group with two proposed methods, *i.e., iBAT*, which detects whether a trajectory is anomalous when it's finished [125], and *iBOAT*, which detects the anomalous events while it's ongoing [17]. However both of them didn't explore what the anomalous behaviors are in nature. This thesis summarizes the types of anomalous passenger delivery behaviors according to their intrinsic properties which direct lead to the detection methods. Then we present an inverted index mechanism based method to improve the recognition efficiency of *iBOAT*. To obtain the digital traces of individual drivers from a two-driver taxi, we for the first time study the work shifting problem. We propose a two-step solution which finds the work shifting agreement first and then detects the daily work shifting events.

With a real-life large-scale GPS dataset collected from a taxi fleet in Hangzhou, China, we detect huge amount of anomalous passenger delivery behaviors and work-shifting events of a community of drivers. We perform analysis on these behaviors from the community perspective. With the 0.44 million anomalous trajectories detected from 7.35 million passenger delivery trips, we perform studies with the intentions to understand the anomalous behaviors, extracting common characteristics of anomalous behaviors, uncovering the motivations behind fraudulent behaviors and investigating the impact of anomalous behaviors on drivers' revenues. After, we analyze the relationship between detour ratio and taxi revenue and find out that high detour tendency doesn't correspond to high revenue. We compare among different taxi groups and find that, to achieve high revenue, taxi drivers need to improve their passenger finding and delivering techniques. So we turn to the study of taxi serving strategies, *i.e.*, the hidden intelligence about taxi serving techniques of a community of drivers based on their digital traces. We propose a novel way to model the passenger finding behaviors as well as passenger delivery behaviors and passenger serving areas of a driver. Then we identify the good and bad strategies by learning from the good drivers, measuring the correlation between strategies and performance, and finding the strategies that can classify driver groups with different performance.

Furthermore, we demonstrate the novel services that enabled by the individual and

community behaviors with a smart taxi systems. It provides various of new services that benefit not only taxi drivers and passengers, but also the taxi companies, the city traffic administrative bureaus and even the public. We illustrate the system working scenarios, interfaces and architecture design. And we evaluate the real time responses and scalability of the system in real life scenarios.

1.2 Thesis Outlines

The outlines of this thesis are organized as follows. Firstly we survey the state-of-art in Chapter 2. Then we introduce the digital footprints collected by pocket-placed mobile phones and GPS-equipped taxis in Chapter 3. In Chapter 4, we present our work in detecting the individual behaviors based on the obtained digital footprints. After, in Chapter 5, we analyze the characteristics of the anomalous passenger delivery behaviors and work shifting behaviors and present how we mine the human intelligence in taxi serving strategies. In Chaper 6, we present the smart taxi system. In the end, we present the conclusions. The main contents of each chapter is summarized as follows.

1.2.1 Digital Footprints Introduction

The digital footprints studied in the thesis include the accelerometer sensing data collected from mobile phones while carried in the costume pockets near the pelvic region and the taxi GPS traces collected from thousands of taxis in Hangzhou, China for one year. They are introduced in the following two separate sections.

1.2.1.1 Accelerometer Digital Footprints

The digital footprints collected by the mobile phones have various scenarios due to the many possible ways of how people carry the mobile phone. And we have to carefully choose the opportunities so that the mobile phones can record the signals of the body movements. This thesis seizes the opportunity when an individual put his mobile phones inside his pockets near the pelvic region to recognize his activities. We build mobile phone interfaces and carefully design experiment scenarios to collect the accelerometer sensing data while people carry the mobile phone in different locations and orientations. Then we give an empirical study about the sensor readings in different scenarios to show the big variance.

1.2.1.2 Taxi GPS Traces

The taxi GPS dataset we obtain is collected in Hangzhou, China from April 2009 to April 2010. We first introduce the city of Hangzhou and how the taxi system works there. Then we give an empirical study about the collected digital traces and further model the

GPS traces with a business cycle that contains four stages, *i.e.*, *Vacant*, *Pick-up*, *Delivery* and *Drop-off*. Besides, we introduce the data problems within the taxi GPS traces with real examples, such as data noises and low sampling problems.

1.2.2 Detecting Individual Behaviors

In this chapter we present how to recognize individual behaviors from one's digital footprints in the obtained two datasets, including detecting one's physical activities from one's accelerometer traces and recognizing anomalous passenger delivery behaviors and work-shifting behaviors from individual taxi's GPS traces.

1.2.2.1 Activity Recognition with Mobile Phone Digital Footprints

Practically, the mobile phones are in natural states inside the pockets around the pelvic region, so they have varying positions and orientations. This thesis presents how to extract features from the raw sensing records and how to perform the classification work on these features to recognize the physical activities. By introducing the acceleration magnitude, which is invariant under different phone orientations, we manage to improve the recognition accuracy about 8% comparing with not introducing the magnitude. The comparison of several data mining methods show that, SVM achieves the highest accuracy. And with a simple feature reduction algorithm, we succeed to reduce the feature dimension to 8 and the least number of feature vectors while maintaining the recognition accuracy.

1.2.2.2 Anomalous Passenger Delivery Behavior Modeling and Detection

We summarize the anomalous passenger delivery behaviors into three types according to their unique properties which lead to different types of solutions. Then we introduce the *iBOAT* method and how we improve the recognition efficiency with an inverted index mechanism. The evaluation result shows that the new method runs at least 5 times faster than *iBOAT*.

1.2.2.3 Work Shifting Behavior Detection

Work shifting events occur in China as most taxis are served by two drivers in order to maximize profit. Observed that the work shifting events normally happen in a fixed location within a short time range of day, which are pre-negoniated by the two drivers, and they take some time for the handover of vehicles, we propose a mapping-based method that firstly detect the work-shifting locations and then find the corresponding work shifting events to separate the digital traces of the two drivers.

1.2.3 Understanding Community Behaviors

With the huge number of anomalous passenger delivery behaviors detected in real life, we perform analysis aiming to answer the following questions. a) What percentage of all trips are anomalous? b) Out of the anomalous trajectories, what percentage of them travel longer distance than necessary? c) What statistical "tendencies" can we discern from the detected anomalous trajectories? d) Do taxi drivers who have a higher tendency to commit fraud have an economical advantage over those who don't? We observe that 1) Over 60% of the anomalous trajectories are "detours" that travel longer distances and time than normal trajectories; 2) The average trip length of drivers with high-detour tendency is 20% longer than that of normal drivers; 3) The length of anomalous sub-trajectories is usually less than a third of the entire trip, and they tend to begin in the first two thirds of the journey; 4) Although longer distance results in a greater taxi fare, a higher tendency to take anomalous detours does not result in higher monthly revenue; 5) Taxis with a higher income usually spend less time finding new passengers and deliver them in faster speed.

We also analyze the spatial and temporal distributions of a large amount of work shifting behaviors. We find that, afternoon work shiftings normally happen between 16:40∼17:20 in non-hot areas, which explains why people feel hard to get taxis in this time period.

We aim to discover both efficient and inefficient taxi serving strategies from the taxi serving behaviors of a driver community. We examine the passenger finding strategies after dropping off passengers, *i.e.* going distant, hunting locally and waiting locally, and before picking up passengers, *i.e.* hunting or waiting, as well as other serving strategies, including passenger serving areas and passenger delivery speeds. To correctly model the strategies after dropping off passengers, we propose a novel method to find the initial intentions right after dropping off passengers in a certain time and location context. By representing the preference of each strategy with a feature, we obtain a feature vector for each driver and a feature document for all the drivers. We first propose a method to learn the most preferred strategies of high revenue drivers. And then we analyze the correlation between each feature and the revenue, to explore the evolving trends and thus to pinpoint the strategies that should be emphasized or weakened by most taxis. In the end, we perform *L1-SVM* and *AdaBoost* to select the most salient features that differentiates taxi performance, and analyze them based on their mutual information with the revenue.

1.2.4 A Smart Taxi System

The perception of individual and community behaviors enables various novel applications of the sensing systems. We illustrate this trend with a smart taxi system with various types of services for three types of users, *i.e.*, passengers, drivers, and system monitors.

The system supports real time tasks like taxi dispatch and anomalous behaviors monitoring as well as other services without prompt response request. We build an easy-to-extend system architecture with the main system components and evaluate the real time responses by "replaying" the collected historical records.

1.3 Thesis Contributions

The contributions of this thesis lie in the following aspects.

1. By mining the accelerometer footprints collected when people place their mobile phone inside the pockets around the pelvic region, we successfully recognize seven daily physical activities. We conduct real life experiments with two Nokia N97 mobile phones to collect the accelerometer readings from 7 volunteers while conducting these activities and prove that, we do able to detect people's physical activities and by introducing the acceleration magnitude into the sensor reading, the recognition accuracy could be improved about 8%. The cross validation results of several data mining methods show that, SVM achieves the best accuracy. With a simple feature reduction mechanism, we successfully obtain a compact model which is much smaller than the original one.
2. We summarize the anomalous passenger delivery behaviors, pinpointing their unique characteristics and the suitable methods to detect them. And then we introduce an inverted index mechanism to replace the trajectory searching in *iBOAT* with index comparisons. The results show that, it improves the computing efficiency at least 5 times.
3. We detect the work shifting events based on the digital footprints of taxis for the first time. By interviewing with several taxi drivers, we find that the two drivers serving one taxi normally have an agreed work shifting location and time period and spend some time for the handover of vehicles. So we extract the waiting locations in the vacant trajectories and map them into grid decomposition. The grids that the taxi stays nearly everyday in a fixed time slot are counted as the possible candidates. After we propose rules to filter out the false candidates and obtain the real work shifting location. And then we detect the work-shifting events and obtain the GPS traces for individual drivers.
4. We detect huge number of anomalous passenger delivery behaviors from the digital traces of a taxi fleet and provide thorough analysis of their characteristics and influences on taxi revenues. We reveal that, these anomalous behaviors normally cost longer traveling distance and more traveling time, which means that most of them are detour events. Even though high detour tendency taxis earn more averagely in

single trips, they are not the top revenue taxis. Top revenue taxis are good at efficient passenger finding and passenger delivering, and don't rely on the fraud behaviors.

5. Understanding the taxi serving strategies aiming to uncover those good and bad techniques is one typical example of uncovering human intelligence from the digital footprints of a community of individuals. We first model the taxi serving behaviors of each driver based on his digital traces from the perspectives of passenger finding, passenger delivering and serving areas. Particularly, we study the passenger finding intentions right after dropping off passenger for the first time and propose a novel method to extract the intentions from the passenger finding trajectories. The taxi serving behaviors of a driver are described as a feature vector, which reveal his preferences over different strategies. Then we study the good and bad strategies from the perspectives of learning from top driver, measuring by the correlation between each strategy and the profitability, and finding the strategies that differentiate the drivers. We present the methods and results accordingly.
6. Furthermore, we present a smart taxi system, which explores the individual and community behaviors mined from the digital traces to support various novel applications. We present the possible users, system architecture and evaluations in real life scenarios.

This thesis presents several studies of detecting the human behaviors with the collected digital traces, analyzing their community characteristics and learning the hidden human intelligence. We hope the illustration of the smart taxi system can inspire novel design of smart sensing systems.

Chapter 2

State-of-the-Art

Contents

We survey the related work from the following two perspectives, detecting individual behaviors from personal sensing digital footprints and understanding community behaviors from community sensing digital footprints. At the same time, we also introduce the applications that are devised from these studies.

2.1 Exploring Individual Digital Footprints

Automated reasoning about human activities is a central goal of ubiquitous computing as well as artificial intelligence. With the advances of sensing and computing, one of the major research fields is to understand human behaviors, goals and intentions, and thus to provide more sophisticated and human-centric services. Generally speaking, such work follows the following steps. Firstly, a low level sensing module is designed to capture various signals about the targeted activities. And then certain features, of which the targeted activities are somehow different, are extracted from the signals. In the end, a module is learned from the features that can tell the activity labels of new testing features [99].

Recent years have been witnessing much research about automatic recognition of human behaviors based on the data collected from mobile and infrastructural sensors. Besides the

vision-based activity recognition work, the majority of studies rely on placing sensor boards on human body to detect the signals of body movements (such as the acceleration measured by accelerometers, and the elevation measured by barometers), or using RFID to detect the objects possibly used in the activity process, and then infer the activities with machine learning or data mining methods. There are also other studies based on the location traces of people (GPS traces and GSM signal traces), which reveal the trips they follow, the time durations they stay at each place, and even what commune tools they take (walking, driving or bicycling). These studies are surveyed as follows.

2.1.1 Sensing-Based Activity Recognition

Current sensing-based activity recognition work mainly relies on three types of pervasive sensing, *i.e.*, ambient sensing, wearable sensing and the combination of the above two [4]. In the following sections, we survey the existing work from these types with consideration about the feasibility in real life scenarios.

2.1.1.1 Ambient Sensing

Ambient sensing refers to the usage of sensors embedded in the environment to perform pervasive sensing tasks. One typical example is the smart home system that captures the environment information and user behaviors and provides services like health monitoring, assistance and information sharing [95]. Earlier examples, such as MIT's PlaceLab [1] [39] and Georgia Tech's "Aware Home" [48], deployed a set of heterogeneous, wired and wireless sensors (with constraints of power supply) in the environment to detect people's behaviors and provide three kinds of services, *i.e.*, health and well-being management, digital media and entertainment, and sustainability.

The sensors in ambient sensing include miniaturized cameras, microphones, RFID, presence and pressures sensors, electricity and water usage detectors. To reduce the adoption difficulties of such systems, a general trend of such system is to use wireless and self-managed sensor nodes [115]. Among these sensors, cameras are common used for activity recognition from vision perspective. For the details of such work, please refer to the surveys [42, 76, 85].

However, in many cases, cameras are intrusive to people's privacies. Thus much research work turns to activity detection based on other types of sensors. In [108] Wilson et al. used many minimally invasive sensors commonly found in home security systems to simultaneously recognize and track activities in a smart home environment. Intille et al. [39] designed a new live-in laboratory "PlaceLab" for studying ubiquitous technologies

1. http://architecture.mit.edu/house_n/placelab.html

in home settings. They deployed various sensors to capture users' activities, such as switch sensors for detecting the open-close events of refrigerator and linen closet, and on-off events of the lighting of a stovetop burner. Logan et al. [68] used a dataset collected from over 900 sensor inputs in "PlaceLab" to recognize dozen of daily living activities and compared the different sensor modalities in detecting various behaviors. In [101] Kasteren et al. studied the activities of a 26-year-old man living a smart apartment.

However, due to the high-cost of building a smart home and high recognition complexity, smart environments that are capable of recognize human behaviors are still far from wide adoption in real life scenarios.

2.1.1.2 Wearable Sensing

Wearable sensing tries to attach various types of sensors in human body to detect the signals that can distinguish various types of activities. The research in this field is promising as people can deploy the sensors in the clothes, belts watches, necklace and mobile phones and thus observe people's behaviors in unobtrusively manner. With the movement signals detected with sensors like the accelerometers, which can be embedded in the devices, many researchers try to recognize people's activities. We introduce the related work according to the ways of the sensor deployment, as it highly influences the feasibility in real life applications.

Activity Recognition with Multiple Sensor Placements

It's intuitive that with more sensors deployed on more locations in the body, we can capture more information about the body movements and thus recognize people's physical activities more precisely. In early research work, Kern et al. [47] fixed accelerometer sensor boards with straps on the major joints on the body, including the locations just above the ankle, the knee, on the hip, wrist and above the elbow and on the shoulder (totally 12 places). By capturing the movement signals of these locations, they could distinguish not only basic postures and movements like sitting, standing and walking, but also activities with the movements of only part of the body, including writing on a whiteboard and typing on a keyboard, and even shaking hands. Bao et al. [6] attached biaxial accelerometers on 4 limb positions plus the right hip to distinguish 20 activities. Laerhoven et al. [53] attached 20 biaial accelerometers along the left leg to distinguish a series of body gestures and movement, such as walking (up/down staris), running, sitting, standing and so on. In [81] Pärkkä et al. placed various types of sensors on different parts of the body to capture not only the movement signals, but also EKG, heart rate, temperature and so on. With such rich information, the authors proposed methods that successfully recognize lie, row, ExBike, sit, stand, run, nordic walk and walk. In [50] Kunze et al. attached MT9 sensor boards on right and left of upper arm, lower leg and knee, the neck and rear hip

to distinguish three types of Tai Chi movements. In [36] Huỳnh et al. attached sensor boards to the right wrist, hip and thigh to recognize a series of low level actives as well as high level ones. In [100] Tapia et al. used five 3-D accelerometers (placed at top of the dominant wrist, side of the dominate ankle, dominate upper arm, dominate upper thigh and dominant hip) and a wireless heart rate monitor to detect not only the physical activities, but also their intensities. Ermes et al. [27] attached various types of sensors all over the body for recognizing activities in both indoor and outdoor environments. Yang et al. [114] designed a body sensing system which consisted of 5 motion sensors attached to the wrists, waist and ankles respectively. The evaluation showed that the recognition accuracy reduced gracefully using smaller set of sensors.

Activity Recognition with Fewer Sensor Placements

Even though multiple sensing positions can recognize sophisticated activities with high recognize accuracy, the sensor deployments are much cumbersome in most of studied scenarios since the sensors were fixed on the corresponding body part with materials like straps, and thus currently still far away from real life adoptions. To make the system easy-to-use, much work turns to use fewer number of sensor placements targeting at more practical use scenarios, and of course, the ambition is reduced to fewer types of activities. In [57] Lester et al. paced three multi-modal sensor boards on the wrist, waist and shoulder with straps to distinguish 8 common physical activities and evaluated the scenarios with just two boards of them. In [104] Ward et al. studied activities about woodwork with microphones and accelerometers mounted on the dominant arm.

Much research work just placed one sensor board in the body to detect the activities. The typical positions are the around the pelvic region [1,20,22,89], shoulder [2,58]and wrist [117]. These types of sensor placements are more practical than the multiple dispositions as the devices can be embedded into commodity stuffs, such as trouser belt, watch and clothes. In an early work, Ravi et al. [89] used just an accelerometer worn near the pelvic region to detect 8 physical activities similar with [57]. Allen et al. [1] used a single triaxial accelerometer mounted on the waist belt to detect three postures and the status change among them. To address the problem of growing rate of sedentary lifestyles, UbiFit Garden [20, 22] used a mobile sensing platform placed on the belt, which was embedded with multiple sensors like accelerometers, barometers, humidity and so on, to infer user's physical activities. The inferred activities were used as inputs of a virtual game in mobile phones to increase people's awareness and interests in promoting physical activities. In [2, 58] the authors also reported the detection of common physical activities with a shoulder mounted multi-sensor board. Besides, Yang et al. [117] also discriminated eight common domestic activities under controlled environment with an accelerometer module mounted on the wrist. The influences of sensor locations in detecting activities were also evaluated in [74].

Activity Recognition with Mobile Phones

Even though the one sensor deployment style is more acceptable in current stage, it's still hard to be promote to mass people, because it's still troublesome to wear such devices (normally they were fixed with bags or clips) and it's costly. The increasingly popular smart mobile phones, like iPhone, Sumsung, Nokia, and HTC, become the ideal platform for practical physical activity monitoring for the public as they are already embedded with relevantly sensing modules. More importantly, they requires no extra cost for hardwares. Much work has been dedicated to this type of research, such as [8,41,51,52,69,71,72,90,116] and ours [97]. Different from the other wearable sensing types which fix certain sensing devices in human body, mobile phone based activity recognition faces new challenges: 1) The sensor location and orientation are not fixed, as users may hold mobile phones in many different possible ways, such as holding in their hands, putting in the pockets (many possible pockets as well) and bags. What's worse, users may change the way of hold mobile phone casually as they want; 2) The mobile phone can be used from many other purposes which influence the recognition of their activities, such as communication and surfing internet; 3) Users won't accept the activity recognition application if it severely slows the system. So the recognition model should not cost too much resources. The reported sensing locations adopted in the work currently include single locations such as the neck [71], front pant leg pocket [52], in a pocket (which pocket is unreported in the paper) [105] and multiple possible locations such as on a table, pants pocket and in the hand in [8], held in the hand, worn on the hip and armband in [69], arm, waist, chest, hand, pocket, and in a bag in [91] and different body locations (details unprovided) in [72]. Conducted at the same time period (year 2010) with most of the work, we also propose to capture the opportunities when people put their mobile phones in their daily costume pockets around the pelvic region [97]. The difference of our study from the previous ones is that, the mobile phones are with many possible location and orientations, which introduces great variances to the model.

2.1.1.3 Combing Ambient and Wearable Sensing

Effective sensor fusion of both ambient and wearable sensing power can take advantage of both modalities to provide novel ways for activity recognition. Most work in this field used RFID bracelet/glove or ear-worn sensors and sensors deployed in the environment to detect the activities. Mcllwraith et al. [75] combined the usage of an ear worn sensor with visual sensors, and proposed a probabilistic sensor correlation framework to identify appropriate set of features for the activity recognition. With the advances of RFID systems, much research turns to use RIFD sensing to detect the objects used and then to infer the conducted physical activities [32, 82, 83, 96, 99, 109, 111]. Normally these studies attached

RFID tags on the objects that used in the targeted activities. And then based on the sequence of object usage, they adopted machine learning or data mining methods to recognize people's activities. Atallah et al. [3] adopted an accelerometer-embedded ear worn activity recognition device combined with wireless ambient sensor in a home environment to identify common activities of daily living.

2.1.2 Other Exploration of Individual Digital Footprints

There are also many studies about people's activities from the digital traces in a long time span, such as the location traces collected with GPS and GSM sensors, location-embedded photos and accelerometer traces. Such digital traces from a long time span potentially record rich information about what an individual has done in which place, and thus could be used to reveal many aspects about one's live, such as the transportation mode, activities and life patterns. The obtained model can also be used to further predict one's future behaviors under various contexts.

Much research focused on inferring the transportation mode, activities, and life patterns from an individual person's GPS digital traces. With the GPS data stream of a traveler, Patterson et al. [25] presented a method that learns a Bayesian model about one's movement, which infers the transportation mode as well as the most likely traveling route simultaneously. Liao et al. [64, 65] found a set of significant places of an individual user in his historical GPS records and then used a Relational Markov Network to recognize the activities in those place, such as working, visiting or dining out. Furthermore, They built a dynamic Bayesian network model to learn and infer transportation routines between the significant places. With such informations, it's possible to obtain a clear understanding of the user's lives, tracking his life patterns and detecting abnormal events, such as taking a wrong bus, which potentially helps cognitively-impaired people to use public transportation safely. Zheng et al. [31, 128] proposed an approach to infer the mobility mode, such as walking, driving and so on based on one's GPS logs. The approach consists of three parts: a change point-based segmentation method, an inference model and a post-processing algorithm based on conditional probability. Perceived that trajectories contain daily whereabouts information of a person, Chen et al. [19] used a clustering method to detect the significant places in a person's life, then abstracted the trajectories and finally extract the movement patterns. With the pattern obtained, the destination of a trajectory and future route can be predicted. Qiu et al. [88] also presented term weighting approaches to mine significant locations from personal location logs. Ye et al. [118] captured an individual's general life style and regularity from his location history. With a history of a driver's destinations and driving behaviors data, Krumm et al. [49] presented a method called *Predestination* that predicts where a driver is going as a trip progresses.

Noticing that a large portion of a typical driver's trips are repeated, Forehlich et al. [45] predicted the route of a driver based on the observation of his past trips. Ziebart et al. [11] presented a method called *PROCAB*, which is used to predict decision at next intersection, route to known destination and the destination given partially traveled route. Monreale et al. [77] proposed a method called *WhereNext* to predict the next location of a moving object, based on the extracted movement pattern called *Trajectory Patterns*. Isaacman et al. [40] inferred important places in people's lives from cellular network data.

There are also many other studies that take advantage of the GPS traces of individual vehicles traveling on the road network with purposes to predict the destination, routings and detecting anomalous behaviors. Edelkamp et al. [26] presented a method to infer a short path to a destination given the current locations and a set of GPS traces. And Cao et al. [13] proposed a method for automatically building a routable road network. Zhang et al. [17, 125] studied the anomalous passenger delivery behaviors of taxis. Anomalous passenger delivery behaviors refers to delivery routes a taxi follows that people normally won't follow to go the destination area from the source. The authors proposed two methods, including *iBAT* and *iBOAT*, to detect the anomalies when the trajectories are finished or while ongoing respectively. However they didn't study the methods from the system point of view. In real practice, the methods will be employed separately on each possible source and destination pairs with testing trajectories. So the computing cost of the algorithms in each single source and destination pair will greatly influence the overall system performance. In this thesis, we propose an inverted index mechanism based method to improve the recognition efficiency. Besides, we collect huge amount of anomalous passenger delivery behaviors from real life GPS traces of thousands of taxis and perform thorough analysis to find the characteristics of the anomalous behaviors from a community perspective.

2.2 Exploring Community Digital Footprints

With the feasibility of collecting digital footprints in community scale, much research attention has been paid to study the community behaviors, revealing the characteristics and their reflections on city dynamics and so on. We survey the related work in groups according to the type of digital footprints them are based on.

2.2.1 Studying the Digital Footprints of Large Group of People

Much research work was conducted based on the GPS traces of large group of people. Firstly, with the location history of different people, the similarities between them can be measured [112, 127]. Xiao et al. [112] modeled the GPS history into a *semantic location history* (SLH), such as *shopping malls* $\rightarrow$ *restaurants* $\rightarrow$ *cinemas* and then measure the

similarity among the SLHs of different people. Zheng et al. [127] presented a GPS-data-driven social networking service called *GeoLife2.0*, where people can share life experiences and connect to each other with their location histories. By mining the location history, it can measure the similarity between different users and recommend friends.

The location histories of a community of people reflect the human mobilities and they can be used to reveal the hidden relationship between different locations (like the correlation in [130]), recommend traveling route [37,130], places and activities for tourist [126]. In [130] the correlated locations are those that usually are visited together in people's trips, and the highly correlated locations can be used to recommended to tour guide, promote sales and design bus routes. Based on multiple user-generated GPS trajectories, Yoon et al. [37] proposed a *Location-Interest Graph* to model typical user's routes in the area. Then given a source, destination and time duration for traveling, it's able to propose an itinerary which is suitable for traveling. In [130], Zheng et al. tried to answer two questions: 1) if we want to do something such as sightseeing or dining in a large city like Beijing, where should we go? (2) If we want to visit a place such as the Bird's Nest in Beijing Olympic park, what can we do there? They modeled the user's location and activity history as a user-location-activity rating tensor and proposed three collaborative filtering based algorithms to handle the data sparseness problem.

2.2.2 Exploring the Digital Footprints from Large Group of Vehicles

There are lots of studies based on the location traces of vehicles. With the wide adoption of GPS devices in vehicles for navigation and other functions like vehicle dispatch (like taxis), there were already huge number of digital traces accumulated from large group of vehicles in real life. The aggregation of such digital traces in large scale potentially reflects many characteristics about human activities and city dynamics. They trigger much research work ranging from urban human mobilities [12, 15, 18, 43, 63, 67, 102], uncovering human behavior patterns [60,66], promoting taxi services [56,60,63,84,86,98,113,123], urban planning [61, 121, 122, 129], estimating regional functions [87, 120] and so on. Following we are going to introduce these research fields separately.

Urban Human Mobility

The research of modeling urban human mobilities based on vehicle traces mainly studies the hotspots of vehicles and taxi pick-up/drop-off behaviors [15, 63, 67, 102], the general human movement patterns [43] and the frequent trajectory patterns [18]. Based on the historical taxi demands, Chang et al. [15] predicted the taxi demand hotspots considering contexts of time, weather and locations. Based on the patterns discovered in the number of passenger pick-ups, Li et al. [63] predicted human mobilities. Viewing vehicles as "sensors", Liu et al. [67] perceived the vehicle crowdedness based on the clustering of the

vehicle mobility. With the taxi GPS traces collected in Lisbon, Portugal, Veloso et al. [102] performed exploratory analysis about the relationships between pick-up and drop-off locations. Based on the moving trajectories of people revealed by the digital traces of 50 taxis during a six-month period, Jiang et al. [43] revealed that the Lévy flight behavior of human mobility patterns is mainly attributed to the underlying street network. Chen et al. [18] proposed efficient methods to discover trajectory patterns.

Smart Navigations

As the passenger delivery trajectories record the drivers' selection of route with the intelligence as a human under different contexts, such as the traffic and road networks, taxi GPS traces are also used to devise smart navigations in urban environment. Letchner et al. [59] proposed a route planning prototype TRIP which incorporates driving preference of the driver and the time-variant traffic conditions which are learned from the GPS traces. Base on the data collected from road speed sensors, Gonzalez et al. [30] found the fastest path by partitioning the map into areas by a road hierarchy, extracting frequently traveled road segments and pre-computing high-benefit paths for each area. Unlike traditional route planning methods that mainly based on the Dijkstra's algorithms, Li et al. [61, 62] incorporated human cognition of the road network which is learned from taxi GPS traces. They divided the road according to the traveling frequencies into frequent roads, secondary frequent roads and seldom roads and then performed route planning by trying to travel through the highest hierarchy roads. In a well known work *T-Drive* [122], Yuan et al. constructed a time-dependent landmark graph, where the landmarks are road segments frequently traversed by taxis, to model the intelligence of taxi drivers and the properties of dynamic road networks. The routing algorithm first finds a rough route on the landmark graph, and then refines it to a detailed route in the road network. In later work [121], Yuan et al. presented a Cloud-mobile architecture system that recommended customized fast driving to an end user considering traffic conditions and driving behaviors. Compared with *T-Drive*, it considers the driver behaviors of the end user and future traffic conditions. It also incorporate the weather conditions obtained on-line and the drivers' knowledge in different regions.

Traffic dynamics

The collective movement of vehicles in a city causes different congestion levels in the road network throughout different time periods of a day. Knowing traffic dynamics is helpful for researches such as promote taxi business and smart navigation. The traffics generally follows regular patterns during the day. Much research work has studied the patterns to better understand the traffic dynamics. With the synergies of hundreds of GPS-embedded taxis which send GPS position to the head quarter once per minute, Schäfer et al. [94] presented a system that reported the real time traffic information. Wen et al. [106]

measured the traffic changes in Beijing around the Olympic game time period based on GPS-equipped taxis. Giannotti et al. [29] found traffic jams by detecting groups of vehicles that move slowly together. Herring et al. [92] estimated the traffic conditions on arterial roads with a Coupled Hidden Markov Model, in which the data from a fleet of 500 taxis in San Francisco, CA, which send GPS data to the server once per minute is used. Yuan et al. [121] inferred the traffic condition at a future time of the landmark graph built from historical data and real-time traffic flow calculated based on recently received taxi trajectories.

Promote Taxi Business

Perceiving the mobility patterns of human as well as taxis, much work has been trying to provide guidances for taxis to efficiently find passengers and vice versa. The studies that tried to help the passenger finding process based on historical GPS traces of large scale taxis evolves from simply recommending hot picking-up areas to sophisticated modelings which consider the competition from other taxis, potential trip length and traffics, and provide optimized passenger finding place and routes.

Recent years have witnessed much research on extracting passenger pick-up patterns and hot areas [15, 56, 63, 86, 102], which can be used for taxi recommendations. As the samples in taxi digital traces record the information including GPS locations, timestamp and passenger status, it's easy to extract the individual passenger pick-up events. And by clustering the locations of these events with *K-means* under the consideration of different time periods of day and day of week, Lee et al. [56] analyzed the pick-up patterns of taxi service in Jeju, Korea and intended to reduce the taxi empty ratio by guiding taxis to these clusters. Chang et al. [15] first gathered the demand request records by filtering with time, location and weather contexts, then clustered these requests into hotspots by *K-means*, *Agglomerative Hierarchical Clustering* and *DBSCAN* and ranked them to provide vacant taxis with good hotspots. Veloso et al. [102] explored the passenger deliveries patterns and passenger finding process, and revealed that in Lisbon, Portugal, for finding passengers, in urban areas taxis normally went to adjacent locations while in suburb areas taxis went to distant locations. Li et al. [63] predicted the passenger pick-up times in the hotspots based on the historical information recorded in the taxi GPS traces and used the results to guide vacant taxis.

Unlike the methods merely focusing on the global hotspots, Powell et al. [86] investigated only the surrounding local areas. They measured the profitability of each area concerning the fare gains of all occupied trips originated from that area, the number of all trips and the cost from current location to the area. Using the knowledge of passenger's mobility patterns and taxi drivers' picking-up/dropping-off behaviors learned from taxi GPS traces, Yuan et al. [123] provided drivers with some locations and the routes to these locations

based on the historical passenger finding probability of the routes and the parking places.

Besides the work based on taxi GPS traces, Takayama et al. [98] used survey information from taxi drivers and proposed promising "waiting/cruising" locations. But the method is prone to human error and inefficient. Yamamoto et al. [113] recommended routing strategies for multiple taxis by mutual exchange of their pathways.

Different from the above perspectives, in this thesis, we intend to learn the taxi serving strategies from the behaviors of taxi drivers based on their digital traces, i.e. to learn the hidden human intelligence during the business process. Since drivers already consider all the influencing factors, we can learn the good and bad behaviors from them considering their behaviors and revenue performance, and then use this knowledge to guide other taxi drivers. The idea is that, we try to study the relations between the behaviors and the resulted revenues, and then reveal good and bad taxi serving strategies to the taxi drivers.

Other research topics

Taxi GPS traces also devise many other novel research topics. Zheng et al. [129] detected flawed urban planning by checking whether the level of connectivity between two areas satisfies the travel demand between them. They looked at actually versus expected distance required to travel between two regions, as well as the expected speed and actual volume of traffic. Their research was based on the trajectories generated by 30, 000 taxis in Beijing (March to May in 2009 and 2010).

Large-scale taxi GPS traces also provide a possible way to measure social functions of city regions. Qi et al. [87] measured the relationship between between social functions of city regions and the pick-up/drop-off characteristics of taxis there. Their results show that the temporal variation of pick-up/drop-off number in a region can depict the social dynamics in that area. Yuan et al. [120] separated the city into disjointed regions according to road structure and inferred the social function in each region using a topic-based inference model. The model is based on human mobility, which is obtained with a taxi GPS dataset, and points of interests located in a region.

Chapter 3

Digital Footprints from Sensing Devices

Contents

3.1 Introduction

According to the wikipedia [1], a digital footprint is a trail left by an entity's interactions in a digital environment, including their usage of TV, mobile phones, internet and world wide web, mobile web and other devices and sensors. Zhang et al. [124] identified three types of digital footprint sources, including the Internet and Web, the static infrastructures and the mobile and wearable sensing devices.

In this dissertation, we focus on the digital footprints collected by pervasive sensing devices. In particular, the digital footprints studied here are the accelerometer sensing data of mobile phones collected when they are carried by people in their pockets of daily costumes and a real life large collection of GPS traces of a large taxi fleet. In this chapter, we give detail introduction about these two data sources.

1. http://en.wikipedia.org/wiki/Digital_footprint

3.2 Mobile Phone Digital Footprints

Recent years have been witnessing an explosion of smart mobile phones. Since the first launch in June 2007, 250 million iPhone units has been sold world wide. And they still lead the trend of smart phones with the new iPhone 5, whose pre-orders were sold out in just one hour. As the major opponent of iOS, Android based mobile phones, such as the Galaxy series of Sumsung and the HTC mobile phones, also draw great attention of the market. Until January 2012, Sumsung has sold more than 50 million Galaxy S and S II mobile phones. In the first quarter of 2011, HTC shipped 9 million smart phones. Beside, Nokia also develops its own smart phones running on the Symbian and Windows Phone 8.

All these smart phones are enhanced with rich sensing capabilities, which potentially leave various types of digital footprints during their usage by people in daily lives. Here list a few of them:

- Proximity sensor: This sensor can determine how close the phones are to users' faces. It helps the mobile phones to turn off the screen automatically when people hold them to the ears to prevent accidental button clicks.
- 3-D Accelerometer sensor: This sensor can detect the acceleration of the device and help to rotate the screen to appropriate views. We seek the opportunity when the mobile phone moves together with human body and thus the accelerometer can measure the acceleration patterns of the corresponding body part.
- 3-D Gyroscope sensor: This sensor measures or maintains the orientation of the device. Together with the accelerometer sensor, it measures the movement in 6 dimensions.
- Microphone: Microphone can not only support the phone calls, but also measure the sounds nearby, such as the noisy level.
- WIFI: WIFI is considered as a sensor because it's able detect the available access points and the wireless signal, which can be used for positioning purpose. A typical use scenario is the digital wall [80], which uses the wireless network to locate the user in indoor environment and provides access control.
- Camera: Modern mobile phones are embedded with powerful cameras. For example, the iPhone 5's camera is 8-megapixel and five-element lens with f/2.4 aperture. It allows people to take pictures conveniently.

The digital footprints collected with these sensors when people carry the mobile phones potentially record many facets about the users. In this thesis, we intend to recognize people's physical activities, for which we choose seven common physical activities that people conduct daily, including stationary, walking, running, bicycling, ascending stairs, descending stairs and driving. Stationary is the status when people are still, including

(a) (b) (c)

Figure 3.1: (a) Pocket locations. For each pocket shown, there is a corresponding one on the left side of the body. (b) Four phone orientations when users put the mobile phone into the right front pocket of the jeans. (c) Coordinate system of the accelerometer sensor in Nokia phones.

standing, sitting and lying down. We seek the opportunities when the mobile phones move together with the human body. As demonstrated in [89], the pelvic region is an ideal accelerometer sensor deployment position for recognizing various physical activities. And also the pockets of normal clothes are designed around this region (i.e. the front and rear pockets of jeans, the front pockets of the coat as shown in Figure 3.1(a)). As revealed by [38], over 60% men get used to putting their mobile phones into their pockets. And thus we are trying to seize the opportunity when people place their mobile phone inside their pocket around the pelvic region to recognize their physical activities.

We choose all the pocket locations shown in Figure 3.1(a) as the potential mobile phone deployment positions. Due to the constraints of pocket shapes, we observed that people usually put the mobile phone into the pockets with a few possible orientations. For example, Figure 3.1(b) shows the scenario when people put the mobile phone inside the jeans pocket. We can see that, normally there are four typical orientations when people put the mobile phones into their pockets. Besides, for the activities people conduct while sitting down, people are not comfortable if they place the mobile phones in the rear pocket of the trousers, as it hurts their butts and may break the phones.

In previous research, the accelerometer has been proven a powerful tool to recognize people's physical activities. In this thesis, we are also trying to use accelerometers in the mobile phones to measure their physical activities. As the sensors are fixed within the mobile phone, their orientations are the same with the mobile phones (Figure 3.1(c) shows the orientation of Nokia phone). So different orientations of the mobile phone cause different sensor readings for the same activity, which introduce great variation for measuring the activities.

We conduct experiment to collect the accelerometer data with Nokia N97 at a sampling

Figure 3.2: Mobile phone interface for labeling the experiment.

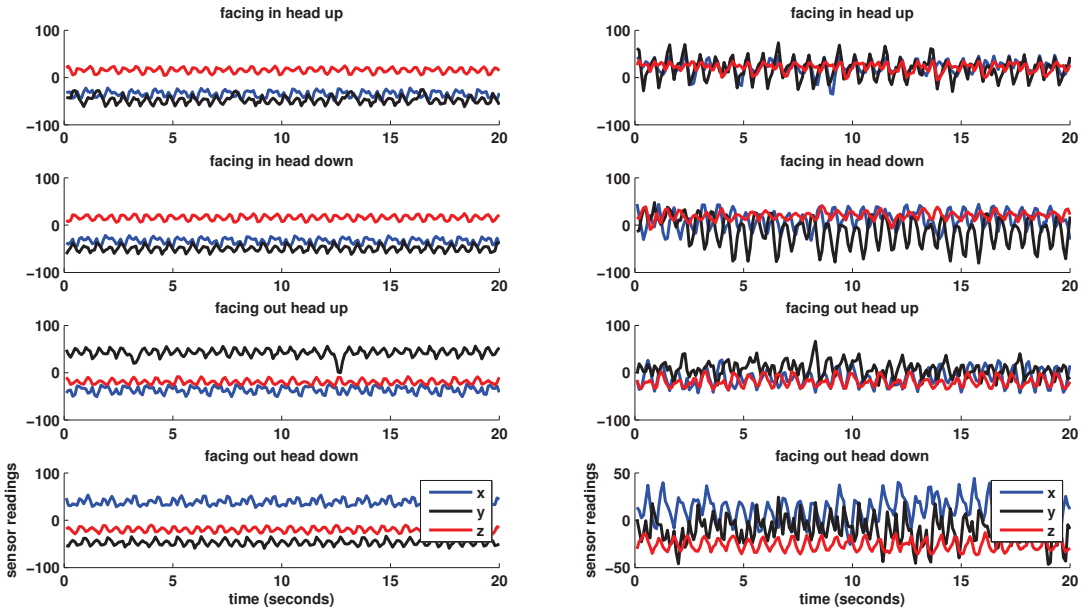
rate of 40Hz for each activity with each possible combinations of location and orientation. To ease the data labeling work, we build a simple touch screen user interface (shown in Figure 3.2.) to label the experiment when launching the application, which allows the users to select what activity they will conduct, which pocket to place in and in which orientation. 7 volunteers (one female and six males) from our campus (Institute TELECOM SudParis) participated in the experiment during a period of three weeks. Before conducting the experiments, they were given an introduction of how to use the application, together with a piece of checklist paper with a list of activities and experiment settings to be conducted with. There were no limitations for the clothes, such as whether to wear tight or loose clothes, or whether to wear a jeans or a pant.

Each time the test subject carried two mobile phones. After launching the sampling application and selecting the correct label for the experiment, the participants put the mobile phones into the targeted pockets according to the chosen settings and conducted the activity for a duration about 5 to 10 minutes. And then they took the mobile phone out and terminated the application. The accumulated accelerometer records during the experiment were saved in a file, whose name is a string with codes for the experiment settings. While residing in the pockets, the mobile phone was in the natural status that could rotate or move. Please also noting that when dealing with the front pockets of the coat, due to the constraints of pocket shapes and user habits, the mobile phone was horizontally facing the body instead of vertically, which was different from the jeans scenarios. Since the first and the last few seconds of the records are the overhead when people put the mobile into the pocket and take it out, they are removed from the official record. In the end, totally we get 48.2 hours training data. The exact time length of each activity is shown in Table 3.1.

Figure 3.3 shows the raw accelerometer readings with different orientations of the mobile phone when placed in the left back pocket of jeans for walking and running. It can be easily seen that, the sensor readings can be quite different when the orientation changes even for

Table 3.1: The sampling time of the activities.

Activity	Station.	Walk	Run	Bicycle	Asc. S.	Des. S.	Drive	Total
Time(Hour)	10.4	9.8	6.3	6.6	4.6	4.0	6.5	48.2



(a) (b)

Figure 3.3: Accelerometer readings when the mobile phone is placed in the left back pocket of jeans with different postures for activities of: (a) walking; (b) running.

the same activity and pocket location. Figure 3.4 reveals the raw accelerometer readings of different activities when the mobile phone is placed in the left front pocket of jeans with the posture of facing in head upward. We can see that they are somehow different.

3.3 Taxi GPS Digital Footprints

The other digital footprints this thesis studies is a large collection of GPS traces from thousands of taxis serving in Hangzhou, China during about one year time period. Before introducing the dataset, we first briefly introduce the general information about the city where our dataset was collected and the taxi service there.

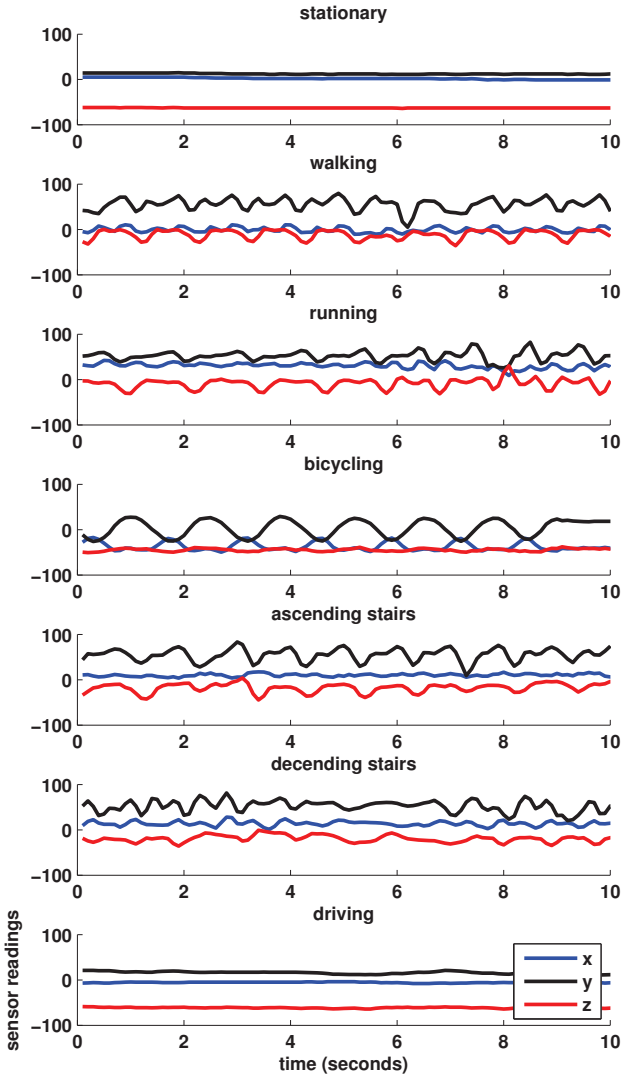


Figure 3.4: Accelerometer readings of different activities when the mobile phone is placed in the left front pocket of jeans with the posture of facing in head upward.

3.3.1 Introduction of Taxi Service in Hangzhou

Hangzhou is the capital and largest city of Zhejiang Province in Estern China and is very close to Shanghai and Suzhou (Figure 3.5(a)(b)). It is a densely populated city with over 8.7 million people dwelling in an area of around 3,000 KM^2 (Figure 3.5(c)). Referred

(a) (b)

(c)

Figure 3.5: Hangzhou in google map. (a) Hangzhou in China; (b) Hangzhou surroundings; (c) The map of Hangzhou city.

as "Heaven on Earth", Hangzhou is famous for its beautiful natural scenery which includes two National Tourist Parks, two National Nature Reserves and several other highly ranked tourist sites. In 2010, it attracted 2.8 million over-night tourists globally and 63 million domestically. To efficiently transport this large population as well as the local dwellers across the city, taxi service plays an important role thanks to the wide availability, convenience and relatively low price. Currently there are about 8,400 continuously operational taxis serving in Hangzhou. People can wave down taxis conveniently or call taxi scheduling center to get a ride quickly. The taxi fare standard when the data was collected worked like this. For the first 3 kilometers, the cost was 10RMB fixed. With additional 7 kilometers, the charge was 2RMB per kilometer. For the part over 10 kilometers, the charge was 3RMB per kilometer.

About 3500 taxis were deployed with GPS sensing devices at the beginning of April 2009 for dispatching. As time passing by, more and more taxis were equipped with such devices and until April 20th, 2010, the number researched about 7600. These GPS sampling devices constantly reported to a central server their GPS coordinates, sampling timestamp, passenger status, speed and orientation via telecommunication network at a rate of about once per minute. The central server received the GPS reports and stored them in a database.

In Hangzhou, all taxis are managed by several big taxi companies who rent taxis to drivers. To obtain high profit, most taxis are served by two drivers, one during the day time and the other during the night time. And thus the service of a taxi is divided into "day work" and "night work". The day driver who serves for "day work" drives a little bit longer than night-work drivers but has to pay more renting fee (210∼230 RMB/day, which is decided by the market) than night-work driver (160∼170 RMB/day).

Each day the two drivers shift work twice, once in the early morning and once in the afternoon. We interviewed with some drivers and found that, the location and time of the work shifting are negotiated beforehand and serve as an agreement for daily shifting. If one driver is late for handing over the taxi to the other driver, he will pay 1RMB/minute to him, or he has to hand over early in the next day in case they have good relationship. So drivers will try to shift the taxis on time except that they can earn more money than the compensation paid. Many night drivers work until 3:00AM in the morning and then park the taxi at the work shifting location and go back home to sleep, while the others serve all night long and then go to the work shifting location to hand over the taxi. The afternoon work shifting happens between 4:30PM and 6:30PM with exact time negotiated by the drivers. Information obtained by interviewing a few taxi drivers reveals that the work shifting locations are normally near petrol filling station, bus stations or driver's home. As taxi drivers may not be close to the shifting location, they normally will plan to go to the work shifting location half an hour earlier, which reduces their ability to serve for requests

Table 3.2: Examples of GPS packets

ID	Longitude	Latitude	Speed	Ori.	Occupa.	Year	Mon.	Day	Hour	Min.	Sec.
9970	120.157762	30.259317	3.7	280	Vacant	2011	11	12	13	2	20
1120	120.258423	30.365834	27	120	occupied	2011	11	12	13	2	50
7869	123.340354	30.289732	18.3	340	Vacant	2011	11	12	13	3	05

with reverse directions and causes difficulties for passengers to find a taxi during this time period. The work shifting event happens within a few minutes (some taxis say it takes about 10 minutes) for checking the petrol tank and handing over.

3.3.2 Taxi GPS Traces Introduction

Some examples of the received reports (they are also referred as *GPS packets* in this thesis) are shown in Table 3.2, which includes taxi ID, Longitude, Latitude, speed, orientation, occupation status and timestamp. With a large taxi fleet, large number of such GPS packets are accumulated each day. For example, during March 2010, we obtained about 441 million GPS packets from the 7600 taxis, which recorded about 7.35 million passenger delivery events happened in that month.

The accumulated records reveal the digital traces of taxis during their business process. In Figure 3.6 we visualize a real life digital trace of a taxi during about 1 hour time period. The markers are the sampling points and their size reflects the waiting time duration with a default size for non-waiting status. The black color trace denotes the passenger delivery processes while the red for passenger finding processes. The changes of status from red to black and from black to red imply passenger pick-up and drop-off events respectively. We can see that, the driver dropped off passengers at location A and drove to location C for finding passengers. After staying there for 10 minutes, he still failed to find them and then decided to go to place B, where it succeeded. It can be seen that, the difference between passenger finding process and passenger delivery process is, normally passenger delivery processes are efficient because the aim is to deliver passengers to their destinations as quick as possible, while passenger finding processes may wander around for finding passengers.

To better understand the business procedure based on the GPS traces, we further separate the digital traces of a taxi into a business cycle that generally contains four stages, *i.e.*, *Vacant*, *Pick-up*, *Delivery*, and *Drop-off* (shown in Figure 3.7). During *Vacant* stage, in order to make the most profit, taxi drivers normally try to find suitable passengers in efficient ways. The influencing factors in this process include the time and driving cost of finding passengers, the potential traveling distance which directly results in the revenue, the traffics and the difficulty of finding the next passenger in the destination of the current

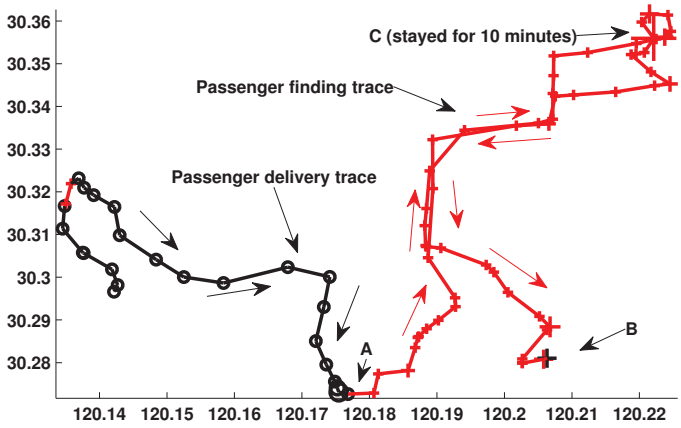


Figure 3.6: Visualization of a real life example of taxi digital trace during about 1 hour time period. Black is for occupied status while red for vacant.

potential one.

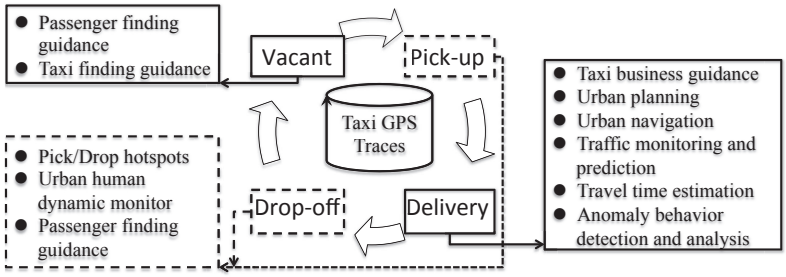


Figure 3.7: Different stages of the taxi business and the possible applications.

The *Pick-up* stage is detected by the state change from vacant to occupied. The aggregation of pick-up events from all taxis reveals the taxi demands in different areas in a city throughout different time periods of a day. Together with the *Drop-off* events which are revealed by the state change from occupied to vacant, we can observe the transport demands among different regions of a city in different time period of day, which is valuable for taxi drivers to find passengers and for city planners to design public transportation and city planning.

After picking up passengers, the *Delivery* process begins. Taxi drivers normally choose the efficient routes to deliver passengers to their destinations. Considering the influences of different traffic situations during a day, there are several possible efficient routes between two areas instead of limiting to the shortest path. A real life example is given in Figure 3.8. It shows all the passenger delivery trips between the source and destination areas. It can

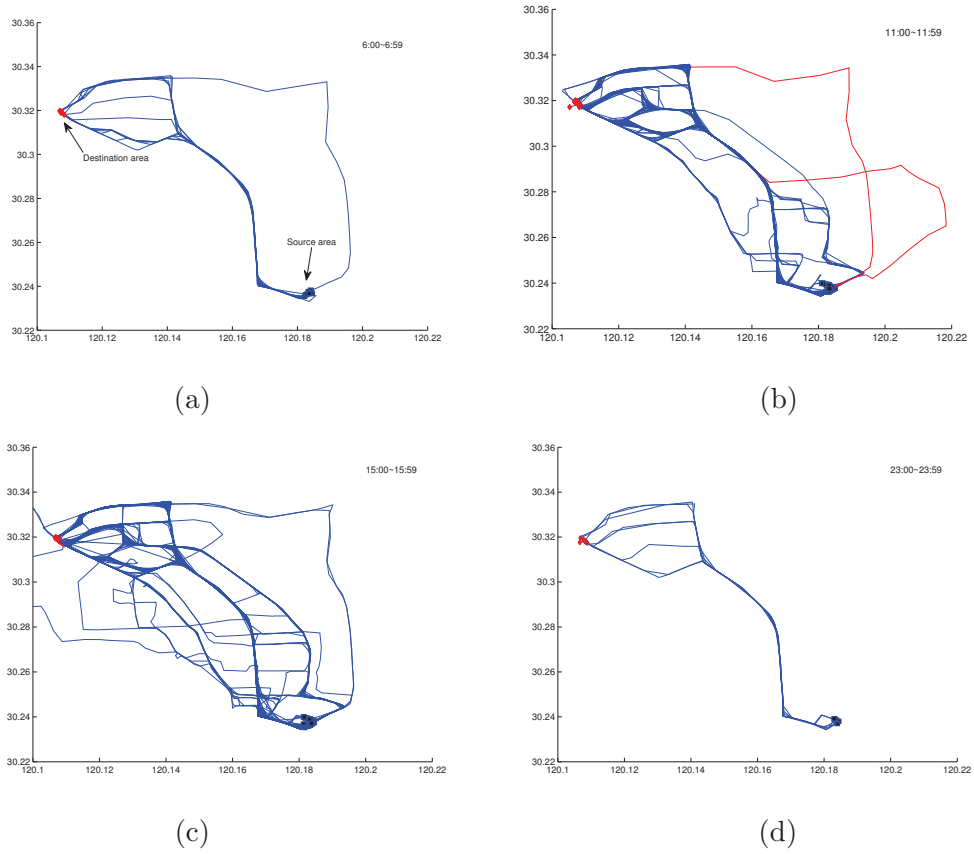


Figure 3.8: The passenger delivery trajectories between two areas in different time periods of a working day. The red lines are two examples of anomalous trajectories.

be easily seen that during rush hours like between 11:00∼11:59 and between 15:00∼15:59, taxi drivers follow many possible routes traveling from the source area to the destination area, while for the non-rush hours like 23:00∼23:59, their choices are reduced.

Taxi drivers may follow some anomalous routes due to reasons like fraudulent driving or sudden-blocked road segments. As revealed in [125], such behaviors are "few" and "different" compared with the majority of the trajectories. For instance, the reds lines in Figure 3.8(b) are two abnormal passenger delivery trajectories. Such anomalous behaviors don't reflect what people normally do under certain situation. In later sections, we are going to introduce the anomalous trajectory detection work conducted in our team and further investigate the characteristics of real life anomalous trajectories.

After dropping off passengers at their destinations, taxis enter the vacant stage again and begin the process of finding passengers. Please note that, *Vacant* stage doesn't necessarily imply passenger finding behaviors. It also include other behaviors, such as shifting

work, having meals, taking rests and so on.

3.3.3 Data Problems

Many problems occur in the real life taxis GPS traces, which may seriously reduce the quality of the sensing data and thus influence the related researches. We summarize the problems together with some of the possible solutions.

- *Data missing.* Data missing usually occurs in the low-sampling-rate situations. For example, a taxi may travel several road segments during the sampling time of one minute. And then we lack of the samples in these passed by road segments.
  Another sever situation is that, sometimes the GPS data is missing in the dataset for several minutes or even hours (e.g., there is about two hour records missing after No. 2 record in Table 3.3). The reasons for this deficiency could be the problems of the GPS sampling devices or a network failure. It results in lacking of information about the movement during the time period. Depending on the addressed problems, the trajectory can either be split into two parts or merely thrown away if infected.
- *Data erroneous.* Occasionally there are errors in some data records, such as the wrong GPS locations or timestamps. For example, in Table 3.3, there is one GPS jump from record No.4 to record No.5, which jumps about 46km far in 75 seconds, and then return back in record No.6. It's high possibly that record No.5 is a GPS device failure as it's impossible for taxis to travel so fast. The samples with this type of erroneous are normally thrown away in real practice. And if necessary, we can "fix" this sample by guessing with the surrounding entries.
- *Multiple drivers.* In many cities, most taxis are served by two drivers to fully take advantage of the vehicle. But there are no direct indication of the drivers in the data. We will provide study about this problem in Chapter 4.
- *Suspicious taxi status.* There are several situations with suspicious taxi status. For example, the status may be set in very short time durations, resulting in very small traveling distance, which is practically suspicious for real passenger delivery. Besides, some taxis may constantly be occupied or vacant for several hours or even several days while moving around, which looks unreasonable in real life. These problems may be caused by the device failure or mal-operation of the driver. We can filter out such deliveries if they are few and not influencing much to the research problem. Besides, we should also consider the time proportion of such improper flag in the overall time of the traces and exclude the taxi with extreme high values.
  Another problem is caused when the driver quickly picks up a passenger right after the drop off events and the device fails to catch such events. Or the drop-off and the following pick-up events are happened in areas without GPS signals. Such events

Table 3.3: Examples of GPS data problems

No.	Taxi ID	Longitude	Latitude	Year	Month	Day	Hour	Minute	Second
1	1111	120.157762	30.259317	2011	11	12	13	1	10
2	1111	120.157762	30.259317	2011	11	12	13	2	20
3	1111	120.258423	30.365834	2011	11	12	15	2	50
4	1111	123.340354	30.289732	2011	11	12	15	3	05
5	1111	123.0	30.0	2011	11	12	15	4	20
6	1111	123.340354	30.289732	2011	11	12	15	5	30

normally happen in the underground or in-building stations of train and airport. In such cases, we fail to capture the drop-off events without awareness of them. This problem is difficult to deal with in most cases. However, for airport, as it's in an isolated area far away from the city, one solution is to insert a vacant entry at the airport and split one passenger delivery trip into two.

– *"Dead" taxis*. Single drivers and some night drivers often stop and go back home to sleep during late night. And sometimes they sleep at some hot locations and wait for passengers to wake them up. Besides, the drivers may need to drive to certain places for having meals, fueling or shifting works. In these cases, their primary targets are not finding passengers (their function for taxi service is "dead"). Such behaviors may influence the observations of the drivers' behaviors. However, such behaviors normally have fixed patterns and thus may be detected. For example, in this thesis, we present the work to detect the work shifting events based on the observation that the shifting events of a taxis normally happen in a fixed location and time.

We filter out those taxis with too much suspicious status or lacking of records for too long time and obtain 6863 taxis to conduct the research.

Chapter 4

Detecting Individual Behaviors

Contents

4.1 Introduction

The digital footprints of individuals provide valuable opportunities to study one's behaviors. In this chapter, we are going to introduce our work in recognizing people's behaviors with the obtained two datasets.

While in the pockets, mobile phones move together with people's bodys and thus their acceleration patterns are similar. So with the obtained accelerometer sensing data, we are possible to find out what kind of physical activities they do. With long term monitoring of their physical activities, we can draw a profile about their physical behaviorial patterns

and tell whether they live sedentary lifestyles or not. However, the mobile phones are in the natural states inside many possible pockets. So their orientations and locations are not fixed, which is different from the previous research. In this work, we conduct experiment to collect the accelerometer data in various locations and orientations. Then we introduce our method to deal with the varying sensing orientations and locations problem.

The GPS traces of taxis record the passenger delivery process and provide opportunities to uncover those anomalous behaviors inside. In this chapter, we introduce what the anomalous passenger delivery behaviors are in their nature and how to improve the efficiency of the existing method to achieve real time responses in real life anomaly detection system. Besides, to obtain the digital traces for individual taxis, we also present how to detect the work shifting events based on the vacant trips.

4.2 Activity Recognition with Pocket Placed Mobile Phones

In this section, we are going to introduce our work in recognizing people's physical activities when they put their mobile phones inside their pockets of daily costumes.

4.2.1 Sensor Data Preprocessing

The embedded triaxial accelerometer inside a mobile phone can continuously sample the experienced accelerations and produce 3-Dimension acceleration readings $A = (a_x, a_y, a_z)$, which are measures of the acceleration experienced in the three orthogonal axes: X-axis, Y-axis and Z-axis. Taking the Nokia mobile phone for example, the coordinate system with respect to the phone body is shown in Figure 3.1(c). When the orientation of the phone body changes, the coordinate system will rotate accordingly and the readings at the three axes will change. Since the acceleration magnitude is a measure for the quantity of acceleration and has no directions, it is insensitive to the orientations of the mobile phones. So it's intuitive that the acceleration magnitude will help to detect the physical activities. As the exact orientation of the acceleration is unknown, to relieve the influences of the phone orientations, we add an the acceleration magnitude to the sensor readings, which thus become 4-Dimension vectors $\bar{A} = (A, |A|) = (a_x, a_y, a_z, |a_x, a_y, a_z|)$.

4.2.2 Feature Extraction

The accelerometer senses discretely the acceleration of the body movement in a creation sampling frequency and generate a sequence of sensor readings $(A^0, A^1, \ldots)$ along the time, where $(A^i = (a_x^i, a_y^i, a_z^i, |a_x^i, a_y^i, a_z^i|))$. To describe the real life physical activities, we model them in separate time slots, with each one a specific activity type. We use a half overlapping window to separate the collected sensing data stream into a number of equal-sized

windows. Within each window $W = (A^0, A^1, \ldots, A^{N-1})$, we extract five types of features, including the *mean* υ, and *variance* ϕ of each sensing dimension, the *correlation* δ among all dimensions, The *DFT energy* η and *Entropy* ϵ in frequency domain and form a feature vector

$$\mathbb{F} = \langle \upsilon, \phi, \delta, \eta, \epsilon \rangle \tag{4.1}$$

. To calculate the frequency domain features, we first perform a discrete Fourier transform (DFT) on each axis in W to get it frequency domain representations. Given a sequence of number $(a^0, a^1, \ldots, a^{N-1})$, DFT transforms it into another sequence of n complex numbers according to the formula:

$$X^k = \sum_{m=0}^{(N-1)} a_m * e^{-i2\pi \frac{k}{N} m}$$

.

The energy feature η is defined as:

$$\eta = \frac{\sum_{k=1}^{N-1} |X^k|^2}{N-1}$$

. And the entropy feature is the normalized information entropy of the DFT component magnitudes excluding the DC component:

$$\epsilon = \sum_{m=1}^{N-1} p^m \log \frac{1}{p^m}$$

, where

$$p^m = \frac{|X^m|}{\sum_{n=1}^{N-1} |t^n|}$$

.

Finally we extract 22 features (4 features for each Mean, Variance, Energy and Frequency-Domain Entropy respectively, and 6 features for Correlation) from each window, which forms the feature vector $\mathbb{F}$ in Equation 4.2. All the feature vectors together build a feature matrix with each column corresponding to one specific feature and each row a feature vector. Each feature vector is labeled with the corresponding activities. Normalization is performed on the extracted feature matrix before training. In a feature matrix T with m vectors, the cth column $\overrightarrow{t_c}(c = 1, 2, \ldots, 22)$ is scaled to [0,1] with the following equation

$$\overrightarrow{t_c i} = \frac{\overrightarrow{t_c i} - Min(\overrightarrow{t_c})}{Max(\overrightarrow{t_c}) - Min(\overrightarrow{t_c})}, i = 1, 2, \ldots, m$$

Practically, we built the feature matrix for different window sizes for two reasons. Firstly, it's intuitive that longer observation time should help to recognize the activities to some extent. So we want obtain the optimized window length to get the best recognition

accuracy. Secondly, within the range of acceptance, larger window length implies that the activity recognition frequency is smaller, which could save the energy consumption but also reduce the sensitivity for activity changes. So choosing an appropriate window length is important for the activity recognition.

4.2.3 SVM-Based Classification Methods

The Support Vector Machine (SVM) [21] is a machine learning method that constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space to solve classification, regression, or other tasks. Whereas the original problem is stated in a finite dimensional space and the classes are not linearly separable in that space, SVM can map this space into a much higher-dimensional space, presumably making the separation easier in that space.

Suppose there are two classes $\mathbb{P}$ and $\mathbb{N}$ to be classified. The training set with n samples are denoted as $\{(\mathbb{F}_i, g_i)\}, n = 1, 2, \ldots, n$, where $\mathbb{F}$ is a feature vector (like in Equation 4.2) and

$$g_i = \begin{cases} 1, & \text{if } \mathbb{F}_i \text{ belongs to } P \\ -1, & \text{if } \mathbb{F}_i \text{ belongs to } N \end{cases}$$

.

A separating plane can be written as

$$W \cdot \mathbb{F} + b = 0. \tag{4.2}$$

Then the margin maximization problem can be represented in dual form as:

$$\begin{aligned} \textbf{Maximize}: \quad & W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} g_i g_j \alpha_i \alpha_j K(\mathbb{F}_i, \mathbb{F}_j) \\ \text{subject to:} \quad & \sum_{i=1}^{n} g_i \alpha_i = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, n \end{aligned} \tag{4.3}$$

. Where α_i is the Lagrange factor and $K(\mathbb{F}_i, \mathbb{F}_j)$ is the kernel function [21]. Then the classification function will be:

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i^* g_j K(\mathbb{F}_i, \mathbb{F}_j) + b^*) \tag{4.4}$$

, where α_i^*, b^* is the optimal solution of Equation 4.3.

4.2.4 Result Analysis

In this section we firstly compare the recognition accuracies of different machine learning algorithms based on the extracted feature matrix as well as the impact of the window length on the performance. Secondly, we evaluate the recognition accuracy improvement with adding acceleration magnitude. Lastly, we perform feature reduction to get a small and compact model, which will cost less computing resources.

Comparison of different algorithms

We adopt WEKA toolkit [35] and LibSVM [14] to perform the classification tasks with the extracted features. We adopt *10-folder cross validation* to get the final accuracy of each trail. All the feature vectors are randomly divided into 10 equal-sized folders. Each time we select one folder as the testing dataset and the rest as the training data set. Then the final accuracies is generated by averaging the results of the 10 experiments. For each experiment, we also perform grid research to obtain the best parameters for each classifiers. The reason why we don't take the training dataset of some people to test the others is because of the population diversity problem. For the details of it, please refer to the section challenge in Section 7.2.1.

The classification accuracy of each algorithms with respect to different window length is revealed in Figure 4.1. It can be seen that SVM performs the best (97.7%) comparing with the rest algorithms. And following is Random Forest [10] (96.5%), whose performance is almost the same as SVM when the tree number exceeds 20. Naive Bayes [44] and RBF Network performs the worst around 70%. When the window length grows from 1 second, the classification accuracy increases for each algorithm. For SVM and Random Forest, it reaches a stable level when the window length is over 6 second. It complies with people's intuition that with longer time observation, the classification accuracy increases and then reaches a stable level. To save the resource consumption, reducing the classification frequency would be acceptable with stable accuracies for some applications. However, the drawback is that the classification granularity would become coarse when the window length increases.

The confusion matrix for the generic SVM model is shown in Table 4.1. We can see that generally speaking, we can classify the activities well. Ascending stairs and descending stairs have bigger chances to mix together and the same situation happens for stationary and driving. It's because intrinsically they are somehow similar.

Improvement with Acceleration Magnitude

We examine the performance improvements of introducing the acceleration magnitude by comparing the recognition accuracies with and without it under different window lengths. As SVM shows the best performance, we choose it for the experiment. The result is shown in Figure 4.2. It can be easily seen that the overall accuracy with acceleration magni-

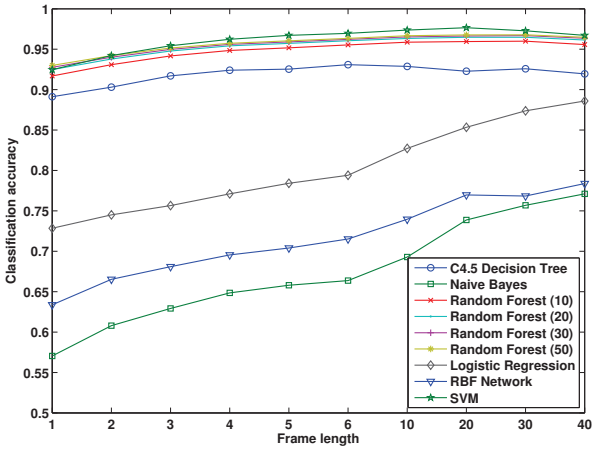


Figure 4.1: The classification accuracy of different algorithms with respect to different window lengths.

Table 4.1: Confusion matrix for the SVM model with window length 6 seconds.

Actual / Predicted	Sta.	Walk.	Run.	Bicy.	Asce.	Desc.	Drive.
Stationary	12776	34	5	75	3	4	129
Walking	88	11923	10	15	91	63	3
Running	60	55	7675	18	21	59	4
Bicycling	104	19	6	7945	23	10	64
Ascending	26	124	4	24	7212	100	4
Descending	72	61	16	48	104	6404	4
Driving	112	7	5	81	6	6	4137

tude outperform those without it. After reaching stable level, introducing the acceleration magnitude improves the accuracy about 8%.

Feature Dimension Contributions and Reduction

To obtain a compact classification model, we evaluate the contribution of each feature attribute to the classification accuracy according to Algorithm 1, which is a loop process with each round excluding the attribute with the least accuracy loss. We choose the window length as 6 seconds and evaluate the contributions with SVM. Figure 4.3 (a) shows how the recognition accuracy varies with the number of feature dimensions left in the evaluation process. It can be seen that when the feature number exceeds 7, the recognition accuracy becomes stable. As the computation cost when predicting with SVM model is directly

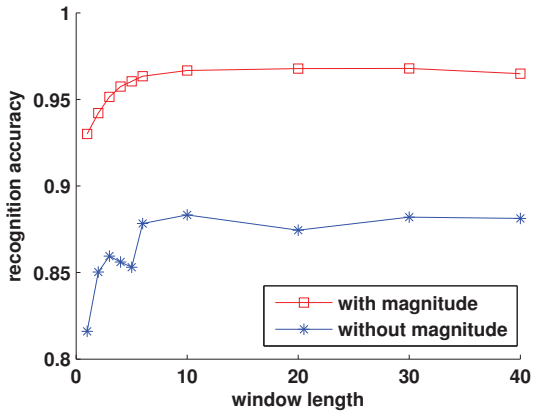
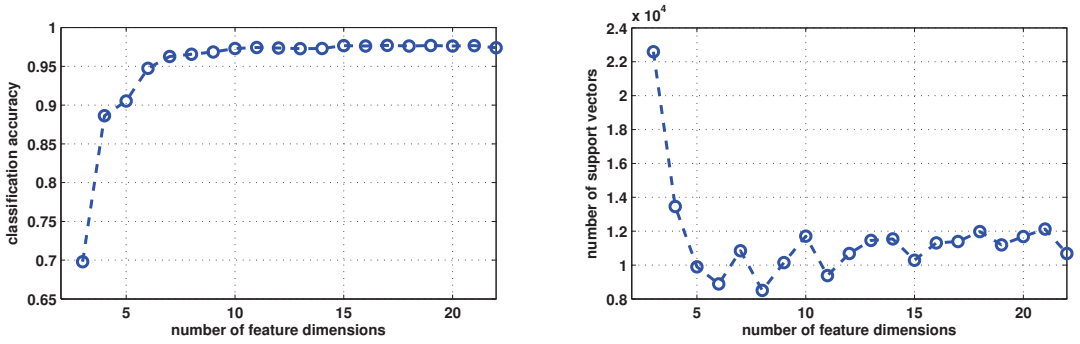


Figure 4.2: The recognition accuracies with and without the acceleration magnitude.



(a) (b)

Figure 4.3: (a) Feature validation result. (b) The number of support vectors versus the number of feature dimensions in the feature contribution evaluation process.

related with the number of support vectors and also the feature dimensions, here we show the number of support vectors with the attribute dimensions left in the evaluation process in Figure 4.3 (b). It's surprising to see that the number of support vectors decreases to the smallest number when the feature dimension is 8 and then increases with less feature attributes. Given observation in Figure 4.3 (a), we can say that these 8 attributes produce a compact model with least computing cost while remaining the recognition accuracy.

4.3 Anomalous Passenger Delivery Behavior Detection

Here we change to the detection of human behaviors from the digital traces of taxis. Generally speaking, anomalous passenger delivery behaviors are those delivery behaviors that aren't often seen in the passenger delivery process. Such anomaly may be caused by

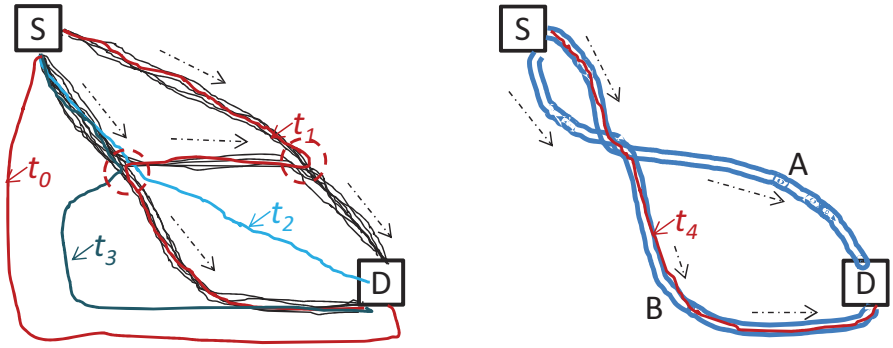
Algorithm 1 Feature Contribution Evaluation

DS is the feature dataset T is the feature dimension of DS
while $1 < T$ **do**
    $Acc_{max} \leftarrow 0$
    for $t = 1 \rightarrow T$ **do**
        $D_t \leftarrow DS$
        Exclude the t^{th} dimension from D_t
        Perform 10-folder cross validation on D_t and get the average accuracy Acc_t
        if $Acc_{max} < Acc_t$ **then**
            $Acc_{max} \leftarrow Acc_t$
            $MinLossD \leftarrow t$
        end if
    end for
    Exclude the $MinLossD^{th}$ dimension from DS
    T is the feature dimension of DS
end while

reasons like fraud driving, blocked or newly built roads or required detours from passengers (for instance, a passenger needs to detour to one place to pick up a friend then go to another place). Currently there are two methods proposed by other colleagues in our team to detect such behaviors inside the taxi GPS traces. One is named *iBAT*, which can detect the anomaly of a complete trajectory, *i.e.*, it can detect whether there are anomalous behaviors inside a trajectory. The other is name *iBOAT*, which can detect the start and end of anomalous passenger delivery behaviors. In this section, we are going to introduce the anomalous passenger delivery behaviors from the perspectives of what they are and what intrinsic properties they have that will help to isolate them. Then we give a brief introduction of the *iBOAT* method and introduce our wok in improving the recognition efficiency, which is critical for adopting such a method in the real time anomalous behavior monitoring of a large taxi fleet in real life.

4.3.1 Anomalous Passenger Delivery Behavior Introduction

As firstly elaborated in [125], the anomalous passenger delivery trips between given a source S and destination D (abbreviated as $\langle S, D \rangle$) pair are those ones that are "few" and "different" from the majority trips. Figure 4.4 illustrates the anomalous scenarios that ever reported in the research. Given the $\langle S, D \rangle$, the majority of the trips are denoted as the black line in Figure 4.4(a) and the two routes A and B in Figure 4.4(b). Besides these normal routes, there are five lines which are not usually followed (t_0, t_1, t_2, t_3, and t_4). In the following sections we are going to analyze these behaviors to give some insights of what they are and how to detect them.



(a) (b)

Figure 4.4: Scenarios of the anomalous passenger delivery trajectories.

Before analyzing the anomalous trajectories, we firstly explain how to deal with the GPS coordinates because they are hard to be dealt with directly as points in a 2 dimensional plane space. We map the GPS coordinates into an finite decomposition of the city, such as the road map or the grid decomposition. In road map decomposition, the GPS coordinates are mapped into road segments with map-matching methods [70,107]. While in grid decomposition, the coordinates are mapped into the grid cells they fall in. We refer to the smallest granularity in a decomposition as the decomposition element (such as a road segment in the road map or a cell in the grid decomposition). Given a set of trajectories T between $\langle S, D\rangle$, we can count the visiting frequency of each decomposition element, *i.e.*, the number of trajectories in T that visit it. If the frequency is high, then the element is traversed commonly and considered as a normal element, otherwise, it's a rare element. In practice, we can decide the "normal" and "rare" attributes by setting a threshold over the visiting frequency.

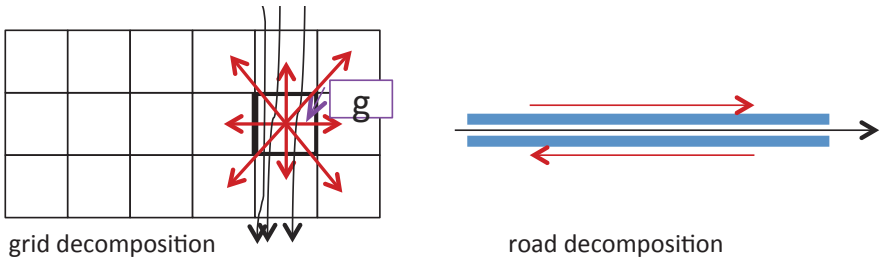


Figure 4.5: Orientations of the grid cell and the road segment.

It can be easily seen that, if a trajectory follows some rare elements (such as t_0, t_2, and t_3), then it's an anomalous trajectory. The rare elements can be safely regarded

as the anomalous segments, which reflect anomalous passenger delivery behaviors. For example, t_0 follows a route that hardly appears in the trajectories traveling from S to D, so t_0 represents an anomalous passenger delivery behavior. Such behaviors can be easily detected by checking whether they follow rare elements.

Another type of anomalous passenger delivery behaviors occurs when the taxi follows normal elements between $\langle S, D\rangle$ but in the reverse direction (such as t_1). Since all the visited segments in t_1 are frequently traversed by other trajectories, this type of anomaly will not be detected if solely based on the visiting frequency of the decomposition elements. However, if we introduce traverse direction of the element when counting the visiting frequency (as shown in Figure 4.5, each grid has 8 directions and each road segment has 2 directions), then instead of calculating the visiting frequency of each decomposition element, we count the visiting frequency of each element in each direction. For example, in grid decomposition shown in Figure 4.5, instead of counting the visiting frequency of a grid, we count the visiting frequencies in all directions. As the anomalous segment in t_1 is rare in the direction, it can be easily detected.

Besides the above two types, there is another rare type of anomalous behaviors ever reported in [17]. As shown in Figure 4.4(b), there are two normal routes A and B between $\langle S, D\rangle$. However, due to some physical constraints, it's very rare that a vehicle can switch from A to B. Then for such an anomalous switch (the red line in Figure 4.4(b)), the anomaly detection with orientation will not work any more as for the switch point, both the directions have high visiting frequency. To detect such anomaly, we need to check the sequence of grids passed through, which describes events like "switch". Then we can adopt the sliding window based method to solve this problem, such as the *iBOAT* method.

4.3.2 Brief Introduction of *iBOAT*

To detect all the aforementioned anomalous behaviors, Chao in our team proposed a method called *iBOAT*, which targets at recognizing the beginning and ending of anomalous behaviors inside a passenger delivery trip. To test the anomaly of a passenger delivery trip t, it first gathers a trajectory set T, in which all the trajectories have the same source and destination area with t. Each trajectory in T is represented as a sequence of grid cells it passes through. To do this, firstly it decomposes the city map into adjacent equal-sized grid cells. Then it maps the sampling points into grids where they fall in. Due to the low sampling problem, the grids mapped may not be consecutive. So it augments them into a sequence of adjacent grids by filling in the gaps between the non-consecutive grids. The testing trajectory t is also mapped into grids where it falls in. But it doesn't need to be augmented.

Given a sub-trajectory t^* which tracks the longest normal grid sequence of a testing tra-

jectory t. *iBOAT* maintains those trajectories from T that contain the same grid sequence of t^* (denoted as T^*). If T^* is rare compared to T (judged by a predefined threshold on $|T^*|/|T|$), it means there are very few trajectories in T that follow the current traveling route, and it reports an anomaly. Then the next grid of t is assigned to t^* and T^* is initialized as T. If T^* is big compared to T, then the next grid in t is added to t^* to extend the sub-trajectory. This process is repeated until the whole trajectory is tested and we can get all the anomaly records in t. Meanwhile, *iBOAT* defines an anomaly score as the traveling length of the anomalous segments.

4.3.3 Improving the Efficiency of *iBOAT*

When testing a sub-trajectory t^*, *iBOAT* searches all trajectories in T^* and maintains those ones that contain t^*. It is a quite time consuming process as it needs to check from the start to the end of each trajectory in T^*. Actually it is a perfect case for using Inverted Index Mechanism to speed up. For example, for searching whether a grid sequence $\langle g_1, g_2 \rangle$ is in a trajectory t, instead of searching it from the start to the end of t, we can easily decide it by checking whether both $pos(t, g_1)$ ($pos(t, g)$ is the position of g in t) and $pos(t, g_2)$ exist (*i.e.*, g_1 and g_2 exist in T) and $pos(t, g_1) < pos(t, g_2)$ (*i.e.*, g_1 is in front of g_2 in T).

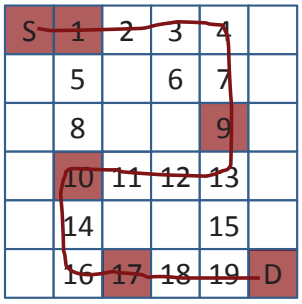
Given a dataset T, we first transform it into an Inverted Index Dataset IID. Let (t, p) denotes an inverted index, which means the pth position in t. Each item g_i : $\{(t_1, pos_1), (t_2, pos_2), \ldots\}$ in IID contains all the inverted indices of g_i in the trajectories belonging to T. We maintain a working indexing set I, initialized to $I = \{(t, 1)|t \in T\}$. The indexing set I maintains all trajectories that are *consistent* with the sub-trajectory t^* currently being examined. An element (t, pos) in I means that trajectory t is consistent with on-going sub-trajectory t^* *up to* position pos. Thus, at the beginning, all trajectories from I begin at position 1.

The process is outlined in Algorithm 2. The indexing set I is initialized in line 1. Lines 2 through 18 iterate when we consider more grid cells of the testing trajectory. In line 3, we fetch the inverted index set G of g in IID. We iterate through the pairs in I in lines 4 through 11, verifying whether each trajectory is still consistent with the new grid cell g: If g is in t (t is occurred in G) and its location is after pos ($pt > pos$), then we change the active position of t in I to be the index of the occurrence of g in t that right *follows* the current active position in t, pos (lines 5–7) ; Otherwise, we remove (t, pos) from I (lines 8–9). Once all the trajectories in I have been checked, we determine whether the resulting size of I is above the allowed threshold (line 12): if so, g is labeled as normal (line 16); if not, g is labelled as anomalous and we reset I (lines 12–14). The process then proceeds with the next received grid cell. Once the trajectory is completed, we can compute an anomalous score using the distance of the anomalous sub-trajectories, as outlined in [17].

Algorithm 2 Improving $iBOAT$ using inverted index mechanism

Input: IID – inverted index dataset generated from trajectory dataset T;
θ – anomaly threshold;
Process:
$I \leftarrow \{(t, 1)|t \in T\}$;
while new grid g of the testing trajectory **do**
  $G \leftarrow IID(g)$;
  for all $(t, pos) \in I$ **do**
    if any (t, pt) exists in G that $pt > pos$ **then**
      $newPos \leftarrow \arg\min_i\{i \geq pos|(t, i) \in IDD(g)\}$;
      update (t, pos) in I to $(t, newPos)$;
    else
      delete (t, pos) from I;
    end if
  end for
  if $|I| < \theta$ **then**
    Label g as anomalous;
    $I \leftarrow \{(t, 1)|t \in T\}$;
  else
    Label g as normal;
  end if
end while
Report the anomalous score of completed trajectory;

To illustrate this process, we go through a simple example. In Figure 4.6, (a) shows the cell IDs in the grid decomposition and a testing trajectory (the red line) with the red cells indicating the position where a GPS sampling happens. (b) lists the historical trajectories and (c) shows the corresponding IID. We depict how the indexing set I changes as the trajectory progresses in (d). All 9 trajectories support the first cell S; this number drops to 8 in cell 1, and notice that the active position in all trajectories increases to 2; upon landing in cell 9 only three trajectories remain, and there are no supporting trajectories at 10 (this means that there are no historical trajectories that include the path $S \rightarrow 1 \rightarrow 9 \rightarrow 10$); at this point the indexing set I is reset to its original state, and on landing on cell 17, there are only 3 trajectories supporting it; then the destination is finally reached. For simplicity we say the anomalous threshold is set at zero. Thus, in this example, only one point was labelled as anomalous: grid cell 10. We can see this process is quite efficient as it can directly pinpoint those trajectories that contain the testing sub-trajectory.



(a) Traversed cells

Grid cell	Contents of I	$\|I\|$
S	$\{(t_1,1),(t_2,1),(t_3,1),(t_4,1),(t_5,1),(t_6,1),(t_7,1),(t_8,1),(t_9,1)\}$	9
1	$\{(t_1,2),(t_2,2),(t_3,2),(t_4,2),(t_5,2),(t_7,2),(t_8,2),(t_9,2)\}$	8
9	$\{(t_2,7),(t_3,6),(t_4,5)\}$	3
10	$\emptyset$	0
17	$\{(t_1,7),(t_5,7),(t_6,6)\}$	3
D	$\{(t_1,10),(t_5,10),(t_6,9)\}$	3

(d) Evolution of indexing set I

$t_1 : S \to 1 \to 5 \to 8 \to 10 \to 14 \to 17 \to 18 \to 19 \to D$
$t_2 : S \to 1 \to 2 \to 3 \to 4 \to 7 \to 9 \to 13 \to 15 \to 19 \to D$
$t_3 : S \to 1 \to 2 \to 3 \to 7 \to 9 \to 13 \to 15 \to 19 \to D$
$t_4 : S \to 1 \to 2 \to 6 \to 9 \to 13 \to 15 \to 19 \to D$
$t_5 : S \to 1 \to 5 \to 8 \to 10 \to 14 \to 17 \to 18 \to 19 \to D$
$t_6 : S \to 5 \to 8 \to 10 \to 14 \to 17 \to 18 \to 19 \to D$
$t_7 : S \to 1 \to 5 \to 8 \to 10 \to 11 \to 12 \to 13 \to 15 \to 19 \to D$
$t_8 : S \to 1 \to 5 \to 8 \to 10 \to 11 \to 12 \to 15 \to 19 \to D$
$t_9 : S \to 1 \to 5 \to 8 \to 11 \to 12 \to 13 \to 15 \to 19 \to D$

(b) Trajectory dataset

($grid$)**S**: $\{(t_1,1),(t_2,1),(t_3,1),(t_4,1),(t_5,1),(t_6,1),(t_7,1),(t_8,1),(t_9,1)\}$; **1**:$\{(t_1,2),(t_2,2),(t_3,2)\ (t_4,2),(t_5,2),(t_7,2),(t_8,2),(t_9,2))\}$; **2**:$\{(t_2,3),(t_3,3),(t_4,3)\}$; **3**:$\{(t_2,4),(t_3,4)\}$; **4**:$\{(t_2,5)\}$; **5**:$\{(t_1,3),(t_5,3),(t_6,2),(t_7,3),(t_8,3),(t_9,3)\}$; **6**:$\{(t_4,4)\}$; **7**:$\{(t_2,6),(t_3,5)\}$; **8**:$\{(t_1,4),(t_5,4),(t_6,3),(t_7,4),(t_8,4),(t_9,4)\}$; **9**:$\{(t_2,7),(t_3,6),(t_4,5)\}$; **10**:$\{(t_1,5),(t_5,5),(t_6,4),(t_7,5),(t_8,5)\}$; **11**:$\{(t_7,6),(t_8,6),(t_9,5)\}$; **12**:$\{(t_7,7),(t_8,7),(t_9,6)\}$; **13**:$\{(t_2,8),(t_3,7),(t_4,6),(t_7,8),(t_8,8),(t_9,7)\}$; **14**:$\{(t_1,6),(t_5,6),(t_6,5)\}$; **15**:$\{(t_2,9),(t_3,8),(t_4,7),(t_7,9),(t_8,8),(t_9,8)\}$; **17**:$\{(t_1,7),(t_5,7),(t_6,6)\}$; **18**:$\{(t_1,8),(t_5,8),(t_6,7)\}$; **19**:$\{(t_1,9),(t_2,10),(t_3,9),(t_4,8),(t_5,9),(t_6,8),(t_7,10),(t_8,9),(t_9,9)\}$; **D**: $\{(t_1,10),(t_2,11),(t_3,10),(t_4,9),(t_5,10),(t_6,9),(t_7,11),(t_8,10),(t_9,10)\}$;

(c) Inverted Index Dataset (IID)

Figure 4.6: An example of *iBOAT* with inverted indexing mechanism. (a) Example trajectory (red line) with mapped grid cells (red squares), blue lines are the grid decomposition of the map; (b) Trajectory dataset from S to D; (c) The corresponding Inverted Index Dataset; (d) Evolution of the indexing set I as the incoming trajectory progresses.

4.3.4 Evaluation Result

We evaluate the performance of Algorithm 2 comparing with the original design in [17] under different historical trajectory dataset size. We choose about 10,000 testing trajectories from 300 $\langle S, D\rangle$ pairs with varying dataset size. For each trajectory, we perform the anomaly detection with both algorithms under the same computing environment and measure the computing time. Then we calculate the ratio between time needed in the original algorithm and time needed by Algorithm 2, and plot the ratio against the size of

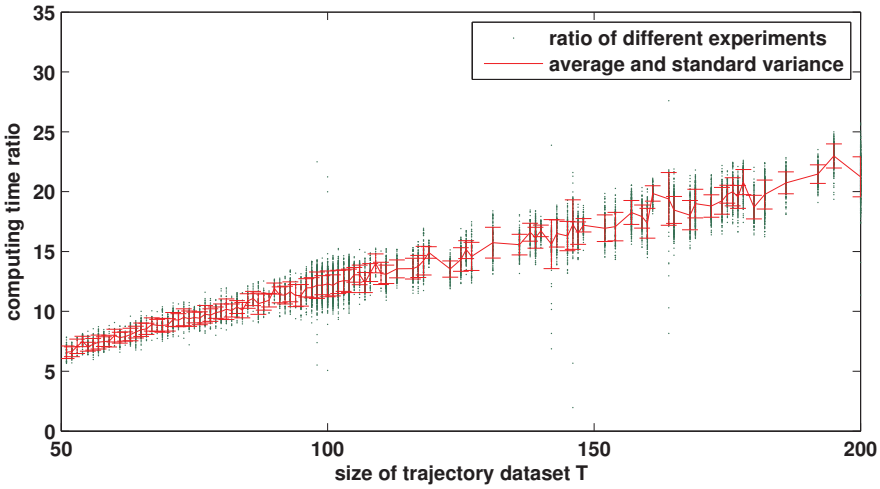


Figure 4.7: Computing time ratio (original/Algorithm 2) versus the size of trajectory set T.

dataset in Figure 4.7.

It can be easily seen that Algorithm 2 is much faster than the original implementation (the average improvements are over 5 times faster). The red line depicts the trend of the mean ratio versus increasing dataset size. It reveals that generally speaking, the ratio increases linearly when the dataset size becomes larger, meaning that the bigger the dataset is, the more advantage Algorithm 2 achieves.

4.4 Work-Shifting Detection

In order to study an individual driver's behaviors based on the GPS traces of a taxi, we have to extract one's personal digital traces first. As the digital traces don't have clear indication about the driver, it's unknown when the work shifting events happen. In this section, based on the spatial-temporal patterns revealed in the traces, we want to find out the work shifting events and separate the digital traces for each individual driver.

4.4.1 Work-Shifting Detection

Based on the observation that work shifting events normally happen in a pre-negotiated place around an agreed time period, we detect the work shifting events with the following two steps. Firstly, we detect the work shifting locations by searching whether there exists a location where a taxi goes routinely in the same time period of every day. Then we identify the daily work shifting events and separate daily taxi GPS traces accordingly.

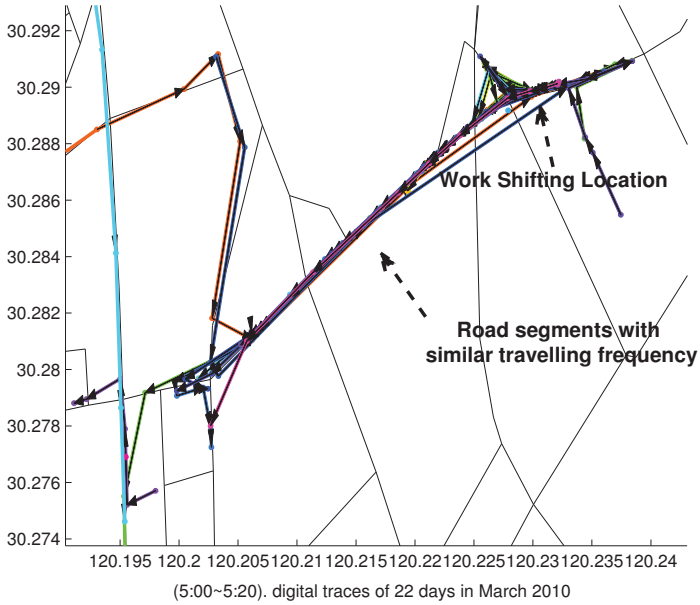


Figure 4.8: Daily work shifting traces. In March 2010, this taxi went to the work shifting location during 5:00∼5:20 for 22 days, and parked there to shift work.

Figure 4.8 shows a real life example of daily work-shifting traces of a taxi in a month. In 22 days of March 2010, the taxi went to the work shifting place within 5:00∼5:20. And by expanding the time slots one hour before and after, we eventually get the visiting records in this place in 31 days. As normally taxis don't have such strong patterns in real life, the visiting frequency becomes an important feature to identify the work shifting location. However, if we merely depend on it, the road segments a taxi passes through for shifting work may also have similar patterns. To filter out those segments, we turn to what happens in the work shifting process. We observe that taxis usually park for about 10 minutes to handle over the vehicles. So the detection of the work shifting location is based on the following three facts:

1. The taxi parks in the work shifting place for a while as drivers need some time to shift the vehicle;
2. For most of the days, the taxi usually visits the work shifting place within a fixed short time period of day while vacant;
3. By expanding the time slot to a larger range, we are able to find the work shifting events which are delayed or in advance.

The process of detecting the work shifting location of a taxi are elaborated as follows.

Algorithm 3 Parking Place Detection

Input: a trajectory Tr, a distance threshold δ, and a time threshold τ
Output: a set of parking locations $\mathbb{P}$

```
1: Initial a GPS sample queue ℚ = null, i = 1, ParkTimeDuration = 0; ℙ = null
2: while i < ||T|| do
3:     j = i + 1; Put Tr(i) in ℚ;
4:     while j ≤ ||T|| do
5:         Add Tr(j) in ℚ;
6:         if RANGE(Q) ≤ δ then
7:             j = j + 1;
8:         else
9:             Delete the last element (Tr(j)) from ℚ;
10:            break;
11:        end if
12:    end while
13:    if ℚ(tail).t − ℚ(head).t > τ then
14:        ℚ contains a parking location; Get the centroid p of ℚ;
15:        ℙ ← p;
16:    end if
17:    i = j; ℚ = null;
18: end while
19: function RANGE(ℚ)
20:    shortestDist = 0;
21:    for i = 1 → ||ℚ|| − 1 do
22:        for j = i + 1 → ||ℚ|| do
23:            if Distance(ℚ(i), ℚ(j)) > shortestDist then
24:                shortestDist = Distance(ℚ(i), ℚ(j));
25:            end if
26:        end for
27:    end for
28:    return shortestDist;
29: end function
```

Firstly we collect all the vacant trajectories and extract the parking locations inside them with Algorithm 3. The general idea of detecting the work shifting location is that, if a taxi stays inside a small place for sufficient long time, we consider it as a parking location. We build a GPS sample queue $\mathbb{Q}$ to maintain the consecutive samples within a short distance range (*i.e.*, the distance among the samples are within a threshold δ). Figure 4.9 gives an example of steps of detecting the parking places. The elements inside each dashed cycle are the GPS samples maintained in $\mathbb{Q}$ in each loop of line 2∼18 in Algorithm 3. As the time span of the samples in Figure 4.9(b) is bigger than τ, we consider it as a waiting location.

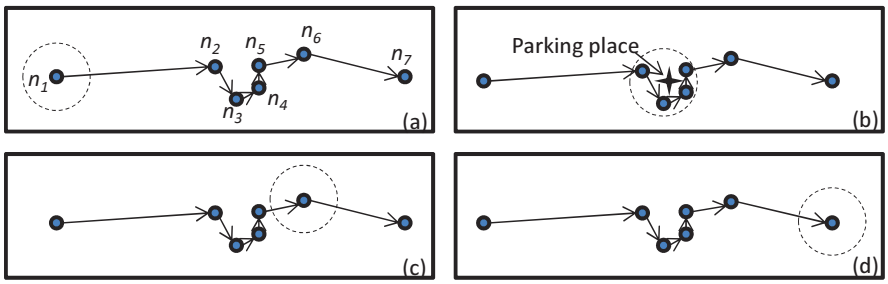


Figure 4.9: Parking place extraction.

In real practice, we choose $\delta = 20$ meters and $\tau = 5$ minutes.

After, we split the city into $50m \times 50m$ large grid cells and separate a day into half-overlapping 20-minute-long time window. In each time window, we record the days that a taxi has parking records in each grid. In case the work shifting place is on the edge of a grid, the shifting events may scatter in different grids. To overcome this problem, when counting the number of days a taxi parked in a grid, we also count the days of the neighboring grids. If the number of the parked days in a grid is bigger than 70% of the total number of days, we will keep it as a possible parking location candidate. Then we expand the time window one hour ahead and one hour later, to include the days corresponding to the delayed and ahead shiftings.

Since we count the neighboring grids, the same work shifting events will make the nearby grids look like possible candidates. Practically we can combine those candidates into one as they are based on the same work shifting logs. In real life, there are some rare occasions that the driver changes the work shifting location temporarily for possible reasons like personal issues or occasionally blocked roads. So we may not be able to get the visiting records of everyday. In practice, for a taxi, if there is one and only one grid achieves 90% of the total number of days within 15:00∼19:00, we say it's the afternoon work shifting place, and if it's within 3:00∼8:00, we say it's the morning work shifting place. In case there are more than one candidate, we consider it as an uncertain (failed) case. With the 6863 taxis, we successfully find the work shifting locations for 4773 taxis. Besides, there are 27 taxis having more than 1 location candidate in the afternoon shifting and 100 taxis in the morning shifting. The rest taxis don't have candidates and exhibit chaos parking patterns, and they are believed to be served only by one driver.

One rare case that may influence the work shifting detection is that, a few number of drivers usually wait at a certain hot place (like the railway station, grand hotels and etc.) for passengers during a certain time period of day. And they may have the same pattern as the working shifting events. However, we observe that in this case the majority of the

corresponding passenger finding records finally succeeded in finding passengers, while the work shifting events normally happen in not-so-hot places and the taxis usually go away right after the shifting while vacant. So we can easily filter such events out based on whether they have passengers after the parking normally.

After obtaining the work shifting location, the digital traces used in counting the visiting days are the work shifting traces, based on which we can easily separate the digital traces for the two drivers. For abrupt shifting location change exceptions aforementioned, for simplicity, we just choose the closest vacant trajectory to work shifting time slots as the work shifting trajectory and separate the digital trace of that day accordingly.

4.5 Conclusion

In this chapter we introduce how to recognize the physical activities with the accelerometer sensing data from people's mobile phones and the anomalous passenger delivery behaviors and work shifting behaviors with the GPS traces of taxis. When dealing with the accelerometer sensing data, before extracting features, we add the acceleration magnitude to the sensor readings, which is invariant to the phone orientations. Experiment shows that the recognition accuracy reaches 97.7% when people put their mobile phones freely into the pockets. Evaluations reveal that SVM achieve the best recognition accuracy and adding the acceleration magnitude could improve the accuracy about 8%. To obtain a compact model, we evaluate the classification contribution of each feature attribute and the number of support vectors in the corresponding model. The obtained result shows that we can reduce the feature dimension to 8 with the minimum number of feature vectors and meanwhile maintain the recognition accuracy.

We reveal the types of the anomalous passenger delivery behaviors based on the characteristics in the GPS traces. By observing from real examples and summarizing from existing work, we conclude 3 types of anomalous behaviors, including following rare decomposition elements, following common elements but in a reverse way and the anomalous switch between normal routes. We elaborate the nature of these anomalous behaviors which leads to different solutions. Then, we improve the efficiency of an existing anomalous trajectory detection method *iBOAT* with an inverted index mechanism at least 5 times faster.

We also detect the work shifting events of a taxi in order to obtain the digital traces of each driver. The work is based on three observations: (1) The taxi parks in the place for a while for drivers to shift the vehicle; (2) For most of the days, the taxi visits the place within a short time slot of day while vacant; (3) By expanding the time slot to a larger range, we are able to find the other work shifting events which are delayed or in advance. We propose a method to detect the waiting locations inside a vacant trajectory first. Then

we select the possible work shifting location candidates and filter out the false candidates by rules. With the 6863 taxis, we successfully find 4773 taxis that satisfy the above criteria. Besides, there are 27 taxis having more than 1 location candidate in the afternoon shifting and 100 taxis in the morning shifting. The rest taxis don't have candidates and exhibit chaos parking patterns, and they are believed to be served only by one driver.

Chapter 5

Understanding Community Behaviors

Contents

Large collections of digital traces from a community of individuals are rich resource for us to find the characteristics about the collective human behaviors and reveal the hidden human intelligence from such behaviors. In this chapter, we present a thorough analysis of the anomalous passenger delivery behaviors from the community point of view. We intend to understand the anomalous passenger delivery behaviors, extracting their common characteristics, uncovering the motivations behind and investigating the impact of anomalous behaviors on drivers' revenues. After, we also investigate the community behaviors of work shifting. We want to see where taxi drivers normally shift work and when and thus give clear indications about the influence of work shifting in taxi service.

After, we try to study the serving strategies of drivers based their digital traces and uncover the good and bad ones which we hope to be able to provide useful guidelines to

taxi drivers. Based on the passenger finding behaviors observed in a community of drivers, whose performance can be acquired from their passenger deliveries, the problem is how to find out the good and bad high level taxi serving strategies, *i.e.*, the strategies that may be helpful or harmful to taxi drivers' revenues. The perspectives of the solutions proposed in this thesis include learning from the behaviors of good drivers, investigating the influence of each strategy separately and seeking the strategies which can differentiate the driver groups with different revenue performance.

Since our study is highly related with taxi drivers' performance, we measure the revenue capability of all drivers and whether their performance is stable, with the intention to see whether the good drivers perform well during all time periods of day. The performance of a taxi driver is influenced by many objective factors like the passenger distributions, passenger destination distributions, potential passenger traveling distance and traffics, and also the subjective factors such as how the driver perceives the objective factors and their preferences and so on. Before studying the behaviors of the drivers, we also provide empirical studies about some of the factors.

5.1 Anomalous Delivery Behavior Analysis

With Algorithm 2, we can collect a large number of anomalous passenger delivery behaviors from all the possible source and destination pairs. We want to analyze the characteristics of such behaviors and provide deep insights about them from the community point of view. Firstly we briefly introduce the configuration used in detecting the anomalous passenger delivery behaviors. As in different time periods of day and day of week, the traffic situations are quite different and thus the normal routes change, we divide a week into working and non-working days, and separate each day into 4 different time slots: night (0:00∼6:59), morning (7:00∼11:59), afternoon (12:00∼16:59) and evening (17:00∼23:59). By combining the type of day and time slots, we get 8 different combinations which we encode as WN (**W**orking day **N**ight), WM (**W**orking day **M**orning), WA (**W**orking day **A**fternoon), WE (**W**orking day **E**vening), NWN (**N**on-**W**orking day **N**ight), NWM (**N**on-**W**orking day **M**orning), NWA (**N**on-**W**orking day **A**fternoon), and NWE (**N**on-**W**orking day **E**vening). We choose the GPS dataset in March 2010. Since the weather is almost the same for the whole month, we simply ignore its influences in the evaluation.

There are about 7.35 million passenger delivery trajectories in the whole month, out of which we successfully obtain 0.44 million anomalous ones. For each anomalous trajectories, we obtain the anomalous sub-trajectories inside with the exact starts and ends. From this large collection of anomalous trajectories, we intend to understand the anomalous behaviors, extracting common characteristics of anomalous behaviours, uncovering

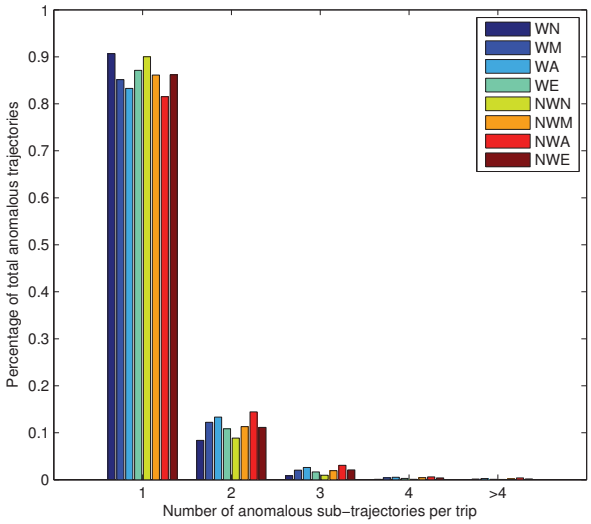


Figure 5.1: Number of anomalous sub-trajectories per trip.

the motivations behind fraudulent behaviours and investigating the impact of anomalous behaviors on drivers' revenues. We thus conduct analysis aiming to answer the following questions.

- What percentage of all trips are anomalous?
- Out of the anomalous trajectories, what percentage of them travel longer distance than necessary?
- What statistical "tendencies" can we discern from the detected anomalous trajectories?
- Do taxi drivers who have a higher tendency to commit fraud have an economical advantage over those who don't?

In this section we aim to discover what are the common characteristics in the anomalous driving behaviors. Although we can't know the exact motives behind anomalous behaviours, these analysis provide clues for these motives, and can potentially increase the detection rate of future anomaly detection methods, as they provide the most pertinent conditions that exist when anomalous behaviours occur.

The first aspect we consider is how many anomalous sub-trajectories occur in one trip. We plot these results for the different time segments in Figure 5.1. And we can see that for all time slots the grand majority of anomalous trips have only one anomalous sub-trajectory.

It is important to also consider at what point during the trip the anomalous behaviour began and ended. We split the trips into thirds and examine where each anomalous sub-trajectory begins and ends. These results are displayed in Table 5.1, and we can see

Table 5.1: Starting and ending positions of anomalous sub-trajectories.

	End			
Start	1st third	2nd third	3rd third	**Total**
1st third	10%	19%	16%	46%
2nd third	N/A	12%	24%	36%
3rd third	N/A	N/A	18%	18%
Total	10%	31%	58%	

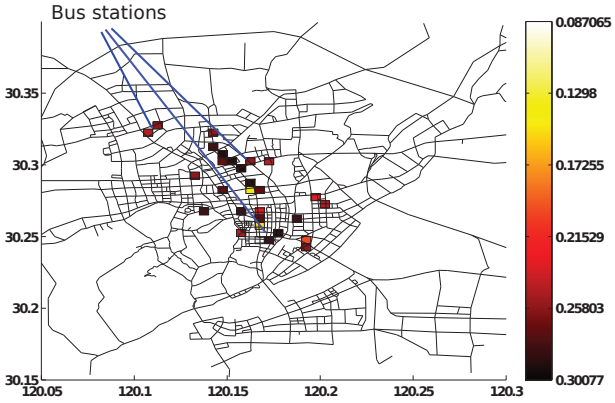


Figure 5.2: Areas where most of the anomalous trips began.

that anomalous sub-trajectories usually don't start and begin in the same third, and most begin in the first or second third. Although this does not clarify the motivations behind the anomalous behaviours, it does suggest that the anomalous behaviours are occuring as a result of a conscious decision, and not by "accident": if drivers had inadvertedly left their intended route, they would generally return to it immediately. In fact, out of all anomalous trajectories, 27% of them remain in an anomalous state until they reach the destination, further reinforcing the belief that these anomalous behaviours are not occurring by "accident".

This is further supported by Figure 5.2, where we display the areas where most of the anomalous trips began. We can see that many of the places are bus stations, where tourists would generally arrive. It is not surprising that they are responsible for a large fraction of the anomalous trajectories. This further confirms our previous claim that anomalous behaviours are conscious decisions.

In Figure 5.3 we display, for varying distances between the source and destination, what proportion of them were anomalous. We can clearly see that the proportion of anomalous trips grows with the trip length, indicating that there is a higher probability of an anomalous trip when the distance between the source and destination is larger. This is consistent with

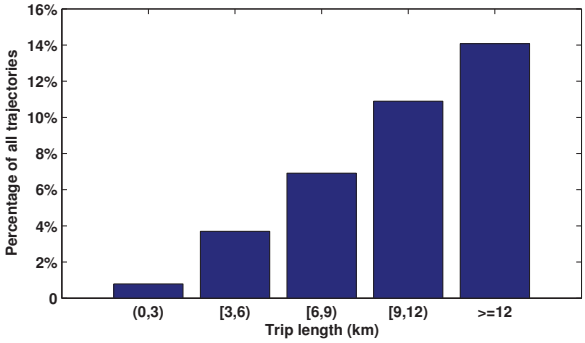


Figure 5.3: The anomalous percentage of all trajectories versus distance between source and destination.

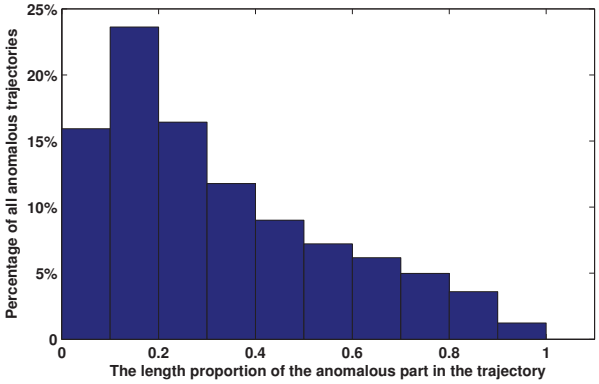


Figure 5.4: Percentage of all anomalous trajectories versus Anomalous length proportion of trajectories.

what one would expect from intuition, as it is easier to take detours in longer trips.

We now examine, out of all the anomalous trips, what proportion of the trips are anomalous (*i.e.* how long are the anomalous sub-trajectories with respect to the whole trip). We display this in Figure 5.4 which demonstrates that even though longer distances have a higher potential for anomalous behaviours, over half of the anomalous trips have anomalous sub-sections that are less than 30% of the trip length.

In most research revolving around detecting anomalous taxi driving behaviours, one is mainly interested in detecting fraudulent activities. We believe that many of these fraudulent trips will take passengers along routes that are much longer than what is considered normal. Given our database of historical trajectories, we can determine the length of the longest normal trip between a source and a destination; we can safely say that an anomalous trip is *detouring* if the trip distance is longer than this maximal distance. For a source-

Table 5.2: Distribution of anomalous trajectories with respect to traveling distance and time.

	Travel time		
Trip length	$[0, minT)$	$[minT, maxT]$	$(maxT, \infty)$
$[0, minD)$	0.0013	0.0137	0.0117
$[minD, maxD]$	0.0062	0.1063	0.0881
$(maxD, \infty)$	0.0045	0.1522	0.6162

destination pair, we denote by $maxD$ and $minD$ the maximal and minimal lengths amongst the normal trips. It may be the case that a longer trip is actually a faster route, placing in doubt whether the driver's actions were fraudulent. We could try to determine $maxT$ and $minT$ for the traveling time taken between two points, but due to varying traffic conditions, these values have a high variability. Because of this, for each source-destination pair, we compute the mean time amongst the normal trajectories, μ_T, as well as the standard deviation σ_T. We then define our boundaries as $maxT = \mu_T + \sigma_T$ and $minT = \mu_T - \sigma_T$. In Table 5.2 we display the distribution of the anomalous trips with respet to these classifications. We can see that over 60% of all anomalous trajectories are taking long detour and longer time, clearly suggesting that fraud is one of the main motivating factors behind anomalous behaviours.

5.1.1 Fraudulent Behaviors versus Revenue

In the last section we observed that most of the anomalous behaviours are due to fraudulent behaviour, and it is safe to say that drivers engage in fraudulent activities for higher revenue, since revenue is based strictly on the distance traveled. It has been previously argued that drivers that take more detours have a higher income [28], in this section we perform a more thorough analysis that demonstrates that the answer is perhaps not as simple as expected.

For each taxi, we compute what we call the *detour ratio* as the total number of anomalous trips divided by the total number of trips. Thus, a higher detour ratio indicates a higher tendency to commit fraud. To estimate the taxi fare for each trip, we use the fare structure of taxi service in Hangzhou to convert the travel distance into income, without considering the waiting time fare compensation (first 3 kilometers, 10 RMB fixed; 2 RMB per kilometer for additional 7 kilometers; over 10 kilometers, 3 RMB per kilometer). In Figure 5.5 we plot the monthly revenue versus the detour ratio (each point is a distinct taxi), as well as the distribution over revenue. We can see that the grand majority of taxis are not prone to detour, and those that have a higher tendency to commit fraud (higher detour ratio) usually have a revenue which is only around average. The correlation be-

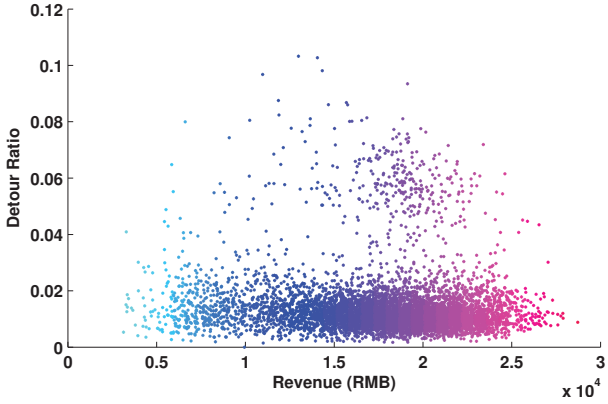


Figure 5.5: Revenues of taxis versus trajectory anomalous rate

tween the taxi revenue and the detour ratio is 3.57%, which is quite small. It indicates that fraudulent behaviour does not necessarily result in higher income, despite popular belief and what was argued in [28].

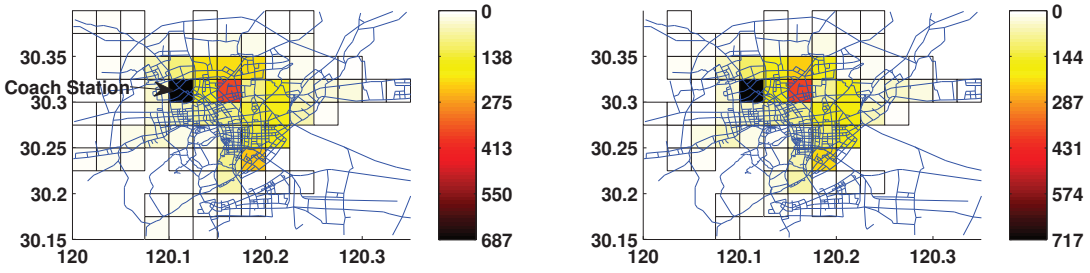
Given that we have revealed that fraudulent behaviours does not necessarily lead to higher income, it is worthwhile trying to discover what differentiates drivers with the highest revenue from drivers with a higher tendency to commit fraud. We compute all the following statistics for the 50 taxis with highest revenue, for the 50 taxis with the highest detour ratio, and for 50 taxis with average revenue. We include this last set of taxis as a baseline to ensure that the differences found between top revenue and top fraudulent drivers are not simply the differences between top and average revenue drivers.

In Table 5.3 we display the average number of passenger delivery trips for each category, the average amount of time taken to find a new passenger, the average speed while serving a passenger, and the average raw revenue per month. We can see that on average, taxis that have a higher tendency to detour serve far fewer taxis per day than top revenue drivers, and even fewer than average income drivers, although the amount of time taken to find a new passenger and the speed during these trips is about the same as average taxis. This means that taxis that have a higher tendency to detour have performance that is almost indistinguishable from average income drivers, *except* for the average number of trips served. This is most likely a consequence of these drivers taking longer trips. Indeed, the average distance travelled by drivers with a higher detour ratio is about 20% than the distance travelled by the other taxis.

Now if we go back to explain the reason why [28] derives a different conclusion from ours, the most probable reason for their differing result is that their study is based on the taxi GPS records from a city where the taxi demand is not as high as in Hangzhou, thus

Table 5.3: Average daily amount of passenger delivery trips and time taken to find new passengers

	trips(#)	Time	Speed(km/hr)	Average Revenue(RMB)
High inc.	50.9	15	25.64	26.8k
Average inc.	37.9	30.25	21.95	20k
High detour	32.3	29.64	21.95	18.6k



(a) (b)

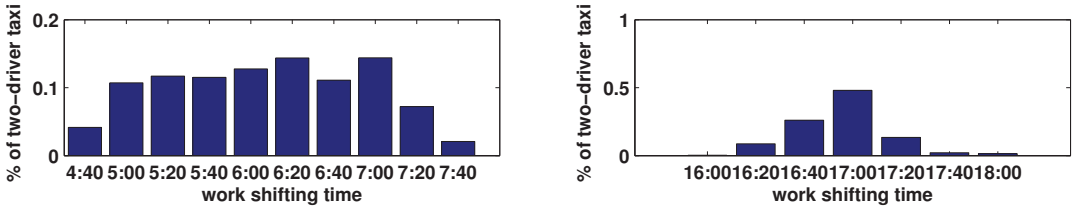
Figure 5.6: Distribution of work shifting places: (a) Morning work shifting; (b) Evening work shifting

taking long detour often means having more income for each trip and as a whole. While in HangZhou, as Taxis are so demanding, thus how to keep the taxi occupied most of the time and deliver the passengers in shortest possible time is more critical for getting an overall high revenue.

5.2 Spatial Temporal Distribution of Work Shifting

We collect the work shifting locations and time slots of all the 4773 taxis and analyze their characteristics in spatial and temporal distribution. We partition the city into $2.4km \times 2.4km$ grid cells and count the number of taxis that shift work in each grid to get the spatial distribution. The work shifting location distribution for morning and evening are shown in Figure 5.6 (a) and (b) respectively. It can be easily seen that both distribution are very similar and most of the morning shifting places locate outside of top hot areas. It's surprising to see that lots of drivers choose to shift near the coach station, perhaps because the drivers can easily take transportation before or after the work shifting.

The shifting time distribution is revealed in Figure 5.7, in which (a) is for the morning shifting time and (b) is for the evening. The morning work shifting time generally distribute between 4:00∼7:20 evenly, while the evening shiftings mainly happen between 16:40∼17:20, which is the one of the major reasons why people feel hard to get taxis around that time



(a) (b)

Figure 5.7: Distribution of work shifting time: (a) Morning work shifting; (b) Evening work shifting

period. In order to alleviate the problem, policies can be adopted to help the drivers to shift work evenly instead of gathering together.

5.3 Taxi Revenue Analysis

Before studying the taxi serving strategies, we first investigate the revenues of all drivers and provide empirical studies about the influencing factors. In particular, we investigate the profitability of the drivers in different time periods of day. Unlike our initial study revealed in [60], which considers the revenues of taxis in a whole day, here we look at a finer granularity (different time periods of day) to check whether the drivers behave well all day or just part of it.

5.3.1 Driver Profitability Analysis

By summing the Euclidean distances between adjacent GPS packets inside a passenger delivery trip, we can estimate its traveling distance and convert it into the raw revenue according to the taxi fare standard (see Section 3.3.1). In order to get rid of the influences of detour behaviors, We replace the revenues of detour trips with the average revenue of the normal trips between the same source and destination area.

The profitability of a driver is measured by its hourly revenue rate, which is calculated by dividing the total revenue by the time duration taken. For example, if a taxi earns 1200RMB during 5:00∼7:00 of 30 days, then the revenue rate is calculated as 1200RMB/30(day)/2(hour)=20RMB/h.

We divide a day into five different time periods, *i.e.*, late night (0:00∼5:59), morning(6:00∼9:59), noon(10:00∼13:59), afternoon (14:00∼17:59) and evening(18:00∼23:59). Since working days may have different patterns from non-working days, we limit our study in working days only in this dissertation. Since drivers may only serve a little period of time in a time slot (for example, a night driver who shifts work at 6:20, may only serve 20 minutes in the

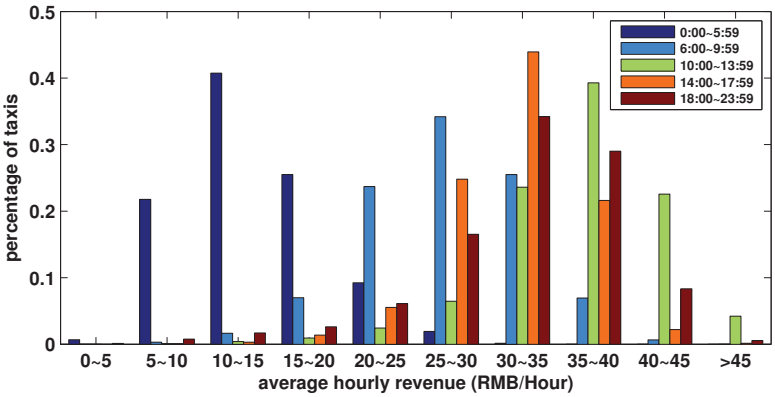


Figure 5.8: Taxi distribution over hourly revenue rate in different time slots of a day.

morning time slot), their behaviors don't reflect the strategies in the whole time period. So we don't count those taxis who serve less than half of the time period in a time slot.

The distribution of taxis over different profitability levels in different time slots is revealed in Figure 5.8. Generally speaking, late night has the worst performance (mostly around 5∼20 RMB/Hour), while noon time has the best (mostly around 30∼45 RMB/Hour), following by the evening and afternoon time period. The distribution of taxis over different performance levels inside each time slot generally follows a normal distribution, which is compliant with the observation revealed in [66].

We investigate the profitability consistency of the drivers over different time periods by ranking all drivers according to their hourly revenue rate. We find that, only 110 night drivers are in the top 500 of both evening and late night time period, and only 67 day drivers in the top 500 of all day time slots (morning, noon and evening), indicating that only less than 22% taxis constantly perform well in night time and only 13.4% in day time. It means only a small fraction of drivers constantly perform well. So when studying the behaviors of taxi drivers, we should investigate them in each time period individually instead of treating all of them as a whole.

5.3.2 Empirical Study about the Influencing Factors

Practically, taxi revenue is influenced by many factors. Here, we are going to give a brief empirical study about two factors, *i.e.*, the passenger delivery times and the taxi demands.

5.3.2.1 Passenger Delivery Times

Intuitively, if taxi drivers deliver more passengers, they earn more money. To validate this intuition, we calculate the correlation coefficient between the number of passenger

Table 5.4: Correlation between number of passenger deliveries and revenue.

	night	morning	noon	afternoon	evening
correlation	0.96	0.84	0.72	0.61	0.86

Table 5.5: Hourly passenger delivery times in each time slot.

	night	morning	noon	afternoon	evening
good	1.3±0.2	2.0±0.2	2.5±0.3	2.1±0.2	2.6±0.2
ordinary	0.7±0.1	1.4±0.2	2.0±0.2	1.8±0.2	2.0±0.2
g/o	1.87	1.4	1.24	1.21	1.31

deliveries (P) and revenues (R) of all drivers. The equation is:

$$corr(P,R) = \frac{\sum_{k=1}^{n}(p_k - \overline{p}) \cdot (r_k - \overline{r})}{\sqrt{\sum_{k=1}^{n}(p_i - \overline{p})^2} \cdot \sqrt{\sum_{k=1}^{n}(r_i - \overline{r})^2}} \tag{5.1}$$

, where p_i and r_i are the total number of passenger deliveries and raw revenue for driver i respectively. We consider each driver as a sample and calculate the correlation of passenger delivery and revenue with all the drivers. The result is shown in Table 5.4 for all the time slots, which most of them are close to 1, indicating that the revenue is highly related with the number of passenger deliveries.

We choose the top 500 drivers as the good drivers and the 500 drivers around the middle income level as the ordinary drivers. We compare the average number of passenger deliveries of them in Table 5.5 for all the time slots. The hourly number of passenger delivery records is in "Mean±Std" format. We can see that good drivers deliver significantly more passengers than ordinary ones, which is at least 21% higher in afternoon and at most 87% higher at night, implying that good drivers are always better at finding more passengers than ordinary ones. So it's always critical for taxi drivers to find passengers efficiently.

5.3.2.2 Passenger Pick-up and Drop-off Hot Areas

One of the main influencing factors is the taxi demands depicted by the passenger hotspots, which have been proposed to direct taxi drivers in finding passengers in previous work [15, 56, 86]. In Figure 5.9 we reveal the top 99 pick-up and drop-off hotspots over different time slots. We can see that the railway stations and the downtown are always hot pick-up and drop-off areas. Residential areas become the drop-off hot spots at night time as many people come back late from the night entertainment hotspots. When moving from downtown to suburb areas, the pick-up rates decreasing greatly. If we look at the hot degrees and rankings of the hotspots among different time periods of day, we will find that they are always changing from time to time. For example, the major entertainment area

annotated in the map becomes less hot in day time than at night time, complying with the fact that people go to work at day time and enjoy the entertainment at night time.

5.4 Understanding Taxi Serving Strategies

5.4.1 Taxi Serving Strategy Introduction

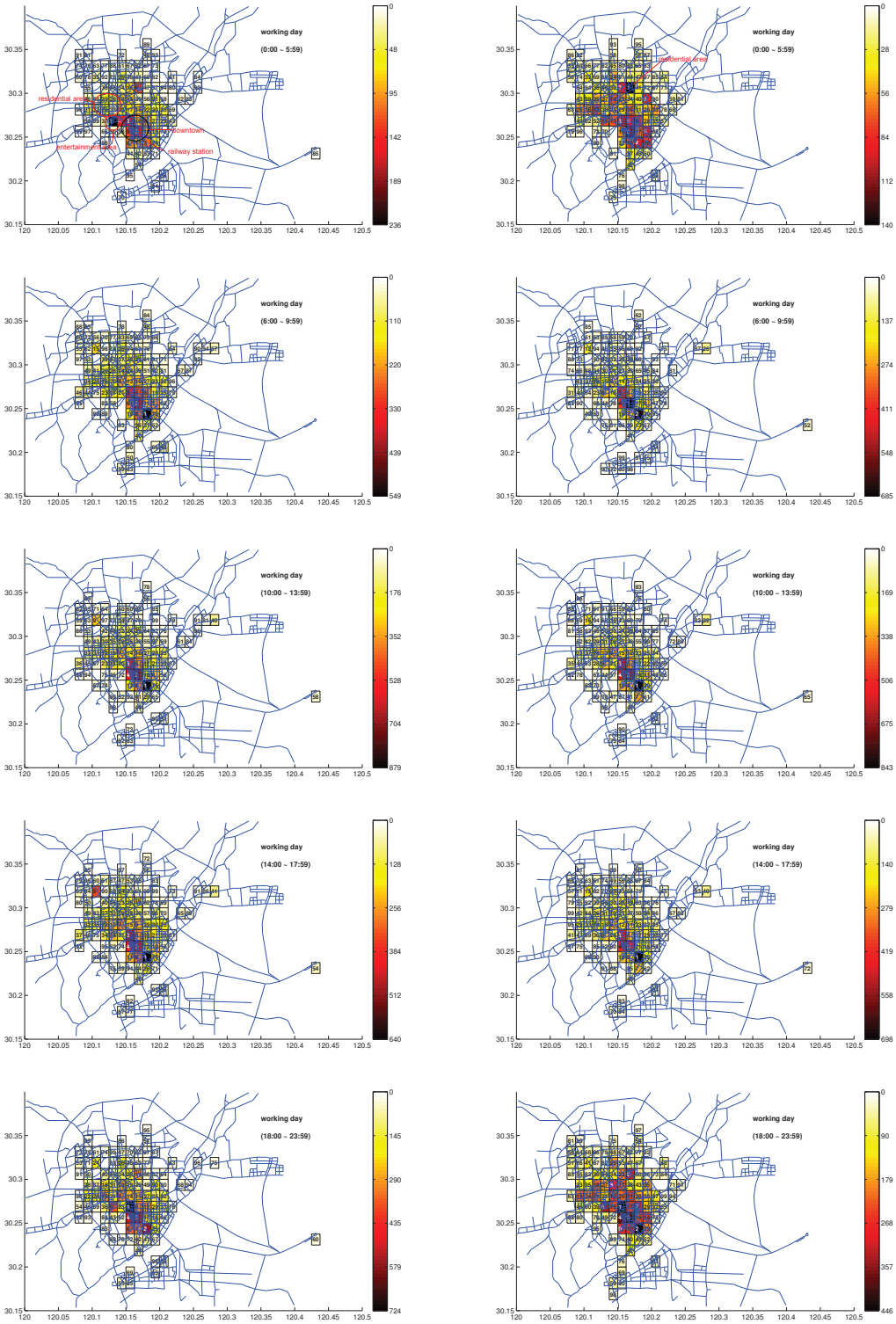
The performance of a driver is influenced by a variety of "time-evolving" objective factors, such as the passenger distributions and their potential travel lengths, and the vacant taxi distributions, which are the demand and supply respectively, and the traffic situations. Meanwhile, it's also influenced by other more "stable" factors such as the hunting cost, which is mainly decided by the cost of fuel and elapsed time, the need to shift work and etc. Moreover, subjectively it's also highly influenced by the driver's awareness of these objective factors (*i.e.* mainly decided by one's experiences), and his reactions accordingly, such as whether hunting nearby, or going to distant hot locations after dropping off passengers, which area to serve, which route to select to avoid the heavy traffic and so on.

Current work about promoting taxi drivers' business mainly focuses on the objective factors, specifically from the demand and supply perspectives [15, 56, 63, 86, 123]. Based on the historical taxi GPS records, much work explored the taxi demands, *i.e*, passenger pick-up patterns and the hot pick-up areas, and recommended to vacant taxis accordingly for finding passengers [15, 56, 63, 86]. Further work considered the competitions introduced by other vacant taxis, and measured the taxi demands by passenger finding possibilities, based on which, a statistically optimized passenger finding location and route can be proposed [123]. Even though the research trend is to incorporate more and more practical influencing factors, it's still hard to acquire a complete and sophisticated model which can perfectly consider all the influences.

Different from the above perspectives, we intend to learn the good and bad taxi serving strategies from the behaviors of a community of taxi drivers revealed in their digital traces, *i.e.* to learn the hidden human intelligence during the business process. Since taxi drivers already consider the influencing factors and their actions lead to different revenue levels, we intend to learn the good and bad behaviors from them by considering the relations between the behaviors and the revenues in a community of drivers, and provide this information back to them for improving their performance.

5.4.2 Taxi Serving Strategy Formulation and Extraction

The drivers' serving strategies studied in this thesis are generally from three perspectives: passenger finding, serving area preferences and passenger delivery techniques. In



pick-up hot map drop-off hot map

Figure 5.9: Top 99 pick-up and drop-off hot areas for each time slot. The number is the hourly number of pick-up/drop-off events in each grid.

particular, in passenger finding strategies we investigate their preference in going distant, hunting locally or waiting locally right after dropping off passengers and their preference in hunting or waiting before picking up passengers. And for passenger serving areas, we study their preference of serving at each area. And for passenger delivery technique, we study their average passenger delivery speed, which implicitly reflects the drivers' ability to choose unobstructed route when delivering passengers. In this section, we introduce how to extract these strategies from the digital traces.

5.4.2.1 Initial Intention Extraction after Dropping Off Passengers

A passenger finding process may consist of several decisions. As shown in Figure 3.6, after dropping off passengers at location A, the taxi starts searching for passengers. It drives to location C, where it waits 10 minutes but still fails to pick up passengers. Then it decides to cruise to place B, where it succeeds in finding passengers. It can be easily observed that this passenger finding trace consists of a sequence of decision making processes, and the final pick-up place is not the result of what the taxi driver decides to do right after the drop-off, otherwise, it would have gone directly to place B. If we simply model the passenger finding strategy of "local" or "distant" as the distance between the current drop-off and the next pick-up locations as in our previous paper [60], then it is "local" as location A and B is close. However, as C is far away from A, so actually the driver wants to go distant right after A. As we focus on the passenger finding strategies right after the drop-off locations, we should investigate what the drivers intend to do right after they drop off passengers (which we call "initial intention path") instead of simply observing the final pick-up places.

Although it's extremely hard to exactly tell the initial mental intention merely based on the GPS traces, we can still have some clues to model their initial intentions. In Figure 3.6 we can observe that, if anomaly occurs from the drop-off location A to the some point B in a passenger finding process, which means that the taxi doesn't follow the efficient way from A to B and there should be a point of interest (POI) (in this case, location C) from A to B that firstly attracts the driver before B. Then B shouldn't be included in the initial intention. Another case is that, a waiting location also is a point of interest that attracts the driver earlier than the POIs after it. So we exclude those passenger segments that are outside of the initial intention and consider the longest normal sub-trajectory starting from the beginning of a passenger finding trace as the initial intention sub-trajectory. Since we can easily cut off the records after the first waiting event to get rid of the POIs after it, the difficult part lies in how to detect the longest normal sub-trajectory in the left part.

A **trajectory** t is a sequence of points $\langle p_1, p_2 \ldots, p_n \rangle$, where p_i is the GPS location (*i.e.* $\langle latitude, longitude \rangle$) sampled during a passenger finding or delivering process. For a passenger finding trajectory tf, whose source area S (where the corresponding drop-off is),

we first cut off the samples after the first waiting position. Then we gather the passenger delivery trajectories that either originate from or pass through S to form a trajectory database TD. Intuitively, as most passenger delivery trajectories follow efficient routes, we can detect the longest normal sub-trajectory of tf from S by checking the longest sub-trajectory from S that is not anomalous compared with TD.

Detecting the anomaly of a trajectory has been well studied in [17,125] when its destination is given. The closest related work is *iBOAT* [17], which, based on a trajectory dataset that shares the same source and destination area, can detect the starting and ending location of the anomaly segment. However, detecting the longest normal sub-trajectory is a problem of detecting the first anomaly starting from S that without a given destination. One intuitive way is assuming that, starting from S, each sampling point is a destination and then we see whether anomaly occurs in it. But since we need to extract the dataset for each destination area and perform the anomaly detection accordingly, it is quite time and resource consuming. In this thesis, we propose a novel method to solve this problem.

Initial Intention Formulation

We denote the ith trajectory in TD as t_i. Like *iBOAT*, for each t_i, we map and augment it into a sequence of adjacent grids where it traverses with the same method used in [125] and denote it as a mapped trajectory $\bar{t}$ ($\bar{t} = \langle g_1, g_2 \ldots, g_n \rangle$). For a mapped testing sub-trajectory $\bar{tf} = \langle g_1, g_2 \ldots, g_m \rangle$, if we can find g_1, $g_2 \ldots$, g_m sequentially in $\bar{t_i}$, then we say t_i complies with tf. We define $hasPath(TD, tf)$ as all the trajectories in TD that comply with tf.

Definition 5.1. Given a sub-trajectory tf whose mapped trajectory is $\bar{tf} = \langle g_1, g_2 \ldots, g_m \rangle$, and a trajectory dataset TD,

$$hasPath(TD, tf) = \left\{ t' \in TD \;\middle|\; \begin{array}{l} \text{(i) } \forall i, 1 \leq i \leq n,\ g_i \in \bar{t'} \\ \text{(ii) } \exists (pos(\bar{t'}, g_1) < pos(\bar{t'}, g_2) < \ldots < pos(\bar{t'}, g_m)) \end{array} \right\} \tag{5.2}$$

We also define $passTraj(TD, g)$ as the trajectories in TD that pass through g.

Definition 5.2. Given a trajectory dataset TD and a grid g,

$$passTraj(TD, g) = \left\{ t' \in TD \,\middle|\, g \in \bar{t'} \right\} \tag{5.3}$$

Now we formalize what it means for a sub-trajectory t to be anomalous with respect to a dataset TD.

Definition 5.3. Given a threshold $0 \leq \beta \leq 1$, a sub-trajectory tf, whose mapped trajectory $\bar{tf}$=$\langle g_1, g_2 \ldots, g_m \rangle$, is β-**anomalous** with respect to a set of trajectories TD if

$$Support(TD, tf) = \frac{|hasPath(TD, tf)|}{|passTraj(TD, g_m)|} < \beta \tag{5.4}$$

The difference between θ-**anomalous** defined in [17] and β-**anomalous** is that, θ-**anomalous** depends on whether the pattern of $\bar{tf}$ is rare in the trajectories of TD, while β-**anomalous** depends on whether the pattern of $\bar{tf}$ is rare in the trajectories from the source to the last grid of $\bar{tf}$ in TD.

The way to detect the initial intention is as following. Starting from g_2 of $\bar{bf}$, if the sub-trajectory $\bar{t'} = \langle g_1, g_2 \rangle$ is not β-**anomalous**, then we include the next grid of $\bar{bf}$ into the sub-trajectory and repeat this process until we find that $\bar{t'} = \langle g_1, g_2 \ldots, g_k \rangle$ is β-**anomalous** or $\bar{t'} = \bar{tf}$, then t' is the initial intention sub-trajectory of tf.

Inverted Index Based Initial Intention Detection

For each grid in the initial intention sub-trajectory except the first, we need to examine every trajectory t in TD to calculate the $hasPath$ and $passTraj$ parameter. It is quite a time and resource consuming process. In this thesis, we propose a novel method which improves the efficiency with trajectory dataset compression and inverted index searching.

The method generally comprises of three parts, trajectory dataset preprocessing, testing trajectory pre-processing and anomaly detection engine (shown in Figure 5.10). For each source area, we accumulate large number of trajectories that either start from or pass through it. As many trajectories are overlapping, which cause great redundancy if we treat them separately. So we compress the whole trajectory dataset into a tree-based architecture TT as illustrated in Figure 5.11. We index each node in TT with (l, p) which l is the node depth and the p is the order in the row (The dashed square in Figure 5.11 is an example). Each traverse from the root to any node (l, p) denotes one unique route and is associate with the number of trajectories that follow it (the black number in the upper left corner of each node in Figure 5.11). We can easily build such a tree with Algorithm 4. For each trajectory t_i in TD, we search from the root of TT (controlled with id). As the first grid of $\bar{t_i}$ is the root, we start from the second grid (controlled with j) and examine whether the jth grid is a child of the idth node of TT. If it is, we move to the child and increase its visiting time vt by 1, otherwise, we create a new node in TT, and add it to the children of the current node, then we set its $vt = 1$ and move to this node (Step 7 $\sim$ 19). We repeat this process on all trajectories in TD and finally obtain the tree. And then we can simply calculate the number of trajectories that pass through a grid g by summing the visiting times of all the nodes in TT whose $nodeID$ is g.

We use an example to illustrate how to calculate the $|hasPath(TD, tf)|$ in Equation 5.4. Let $\bar{tf} = \langle g_1, g_4, g_5 \rangle$, we first find the offsprings of g_1 whose nodeID is g_4 (two red nodes in Figure 5.11) and record their indices. Then we can simply add the vts of these nodes to caluate $|hasPath(TD, \langle g_1, g_4 \rangle)|$=100+160=260. After we search the offsprings of these indices and find those nodes whose nodeID is g_5 (two blue nodes in Figure 5.11). Then $|hasPath(TD, \langle g_1, g_4, g_5 \rangle)|$=10+10=20.

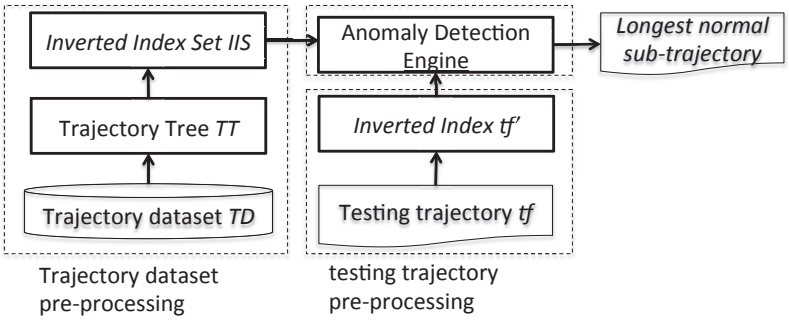


Figure 5.10: Longest normal sub-trajectory detection method.

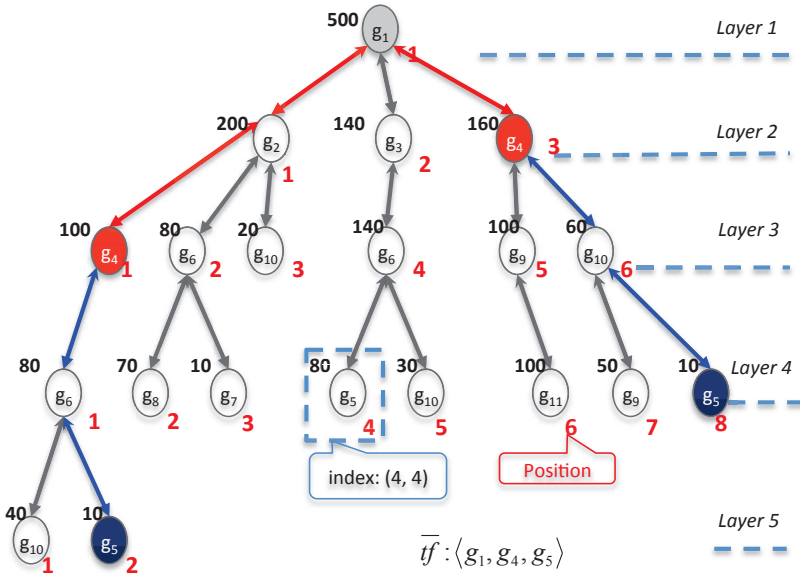


Figure 5.11: Tree representation of the trajectory dataset.

In each round of the aforementioned steps, we need to search all the offsprings of the current indices, which is also quite time consuming. To fasten the searching process, we devise an inverted index mechanism based method inspired by the following idea. If node n_2 is an offspring of n_1, then $n_2.layer > n_1.layer$ and $0 < n_2.vt <= n_1.vt$. If we can define some function $offset(n)$ of the node n, so that when $n_2.layer > n_1.layer$, if n_2 is an offspring of n_1, then

$$offset(n_1) < n_2.vt + offset(n_2) <= offset(n_1) + n_1.vt \tag{5.5}$$

will hold, otherwise, it doesn't hold. Then we can track all the occurrences of each grids in TT and replace the aforementioned searching process with a simple comparison process

Algorithm 4 Converting a trajectory dataset into a tree

Input: TD – a set of trajectories
Output: TT – a trajectory dataset tree represented in a structured array

Process:
$TT \leftarrow$ structured array in the format of $\langle nodeID, vt, father, children\rangle$, $children$ attribute records all the indices of the children nodes;
TT tail index $TID \leftarrow 0$
for $i = 1 : length(TD)$ **do**
    Convert t_i in TD into mapped trajectory $\bar{t_i}(\langle g_1, g_2 \ldots, g_m\rangle)$;
    $id \leftarrow 1$;
    for $j = 2 : m$ **do**
        if $Tree(id).children$ consists of g_j **then**
            $id \leftarrow$ the index of g_j in $Tree(id).children$;
            $Tree(id).vt \leftarrow Tree(id).vt + 1$;
        else
            $TID \leftarrow TID + 1$;
            $TT(TID).nodeID \leftarrow g_j$;
            $TT(TID).father \leftarrow id$;
            $TT(TID).vt \leftarrow 1$;
            $TT(TID).children \leftarrow null$;
            $TT(id).children \leftarrow [TT(id).children\ TID]$;
            $id \leftarrow TID$;
        end if
    end for
end for

of all the occurrences of adjacent grids in $\bar{tf}$.

We find one feasible definition of $offset(n)$. When the order of the nodes in each layer is fixed (numbered from left to right as the red numbers in Figure 5.11), $offset(n)$ equals to the sum of the visiting times vt of all left siblings of the nodes in the traverse from the root to n.

$$offset(n) = \sum n_s.vt, n_s \text{ is the left sibling of a node in the traverse from the root to } n. \tag{5.6}$$

For example, $offset((4,4)) = 200$ as the only left sibling exists in the traverse from the root to $(4,4)$ is $(2,1)$, whose vt is 200. And $offset((3,3)) = 100 + 80 = 180$, as the left siblings are $(3,1)$ and $(3,2)$, whose vts are 100 and 80 respectively.

Proof. For two nodes n_1 and n_2 in TT and $n_2.layer > n_1.layer$, if n_2 is an offspring of n_1, then

$$offset(n_2) + n_2.vt = offset(n_1) + \sum n.vt + n_2.vt \geq offset(n_1)$$

, where node n is the left sibling of the nodes in the traverse from n_1 to n_2. And also since all the trajectories reaching n_1's offsprings pass through n_1, then

$$n_1.vt \geq \sum n.vt + n_2.vt$$

, and thus

$$n_1.vt + offset(n_1) \geq offset(n_1) + \sum n.vt + n_2.vt = offset(n_2) + n_2.vt.$$

Then

$$offset(n_1) < n_2.vt + offset(n_2) <= offset(n_1) + n_1.vt$$

holds.

If n_2 is not an offspring of n_1, then if n_1 is an offspring of the left siblings of the nodes in the traverse from the root to n_2, then

$$offset(n_2) > offset(n_1) + n_1.vt$$

; otherwise n_1 is an offspring of the right siblings of the nodes in the traverse from the root to n_2, then

$$offset(n_2) + n_2.vt < offset(n_1)$$

. □

For each node n in $Tree$, we design an inverted index as

$$\langle l_n, offset_n, vt_n \rangle$$

, in which l, $offset$ and vt are its layer number, offset obtained from the offset function and visiting times respectively. Then for two nodes n and m, if

$$l_n < l_m \tag{5.7}$$

and

$$offset_n < offset_m + vt_m <= offset_n + vt_n \tag{5.8}$$

, then n is an offspring of m, and otherwise it isn't.

We maintain an **I**nverted **I**ndex **S**et *IIS* which contains the inverted indices of all nodes in TT indexed by the grids. For example, the *IIS* of Figure5.11 is shown as in Table 5.6. Then we can replace the tree searching process in Algorithm 4 with the above comparison process. Then the algorithm becomes to be Algorithm 5. *curSet* maintains the current indices which correspond to the indices in TT, to which the routes from the root support the sub-trajectory. When testing a new grid, we update *curSet* with the new locations that

grid	inverted indices	vt
g_1	$\{\langle 1,0,500\rangle\}$	500
g_2	$\{\langle 2,0,200\rangle\}$	200
g_3	$\{\langle 2,200,140\rangle\}$	140
g_4	$\{\langle 2,340,160\rangle,\langle 3,0,100\rangle\}$	260
g_5	$\{\langle 4,200,80\rangle,\langle 4,490,10\rangle,\langle 5,40,10\rangle\}$	100
g_6	$\{\langle 3,100,80\rangle,\langle 3,200,140\rangle,\langle 4,0,80\rangle\}$	300
g_7	$\{\langle 4,170,10\rangle\}$	10
g_8	$\{\langle 4,100,70\rangle\}$	70
g_9	$\{\langle 3,340,100\rangle,\langle 4,440,50\rangle\}$	150
g_{10}	$\{\langle 3,180,20\rangle,\langle 3,440,60\rangle,\langle 4,280,30\rangle,\langle 5,0,40\rangle\}$	150
g_{11}	$\{\langle 4,340,100\rangle\}$	100

Table 5.6: *IIS* of Figure 5.11

Algorithm 5 Longest normal sub-trajectory detection based on *IIS*

Input: *IIS* – inverted index set, $IIS(g)$ is the invert indices for grid g;
tf – A testing trajectory; β – Threshold for anomaly detection
Output: tln – longest normal sub-trajectory
Process:
if tf exists waiting for passengers **then**
    Get rid of the records after waiting;
end if
Get the mapped trajectory $\bar{tf}$ ($\bar{tf} = \langle g_1, g_2 \ldots, g_m\rangle$);
$curSet \leftarrow IIS(g_1)$; $SEQ \leftarrow \langle g_1\rangle$;
for i=2:m **do**
    $candSet \leftarrow IIS(g_i)$; $tempSet \leftarrow \{\}$;
    for each element e_1 in $CandSet$ **do**
        for each element e_2 in $curSet$ **do**
            if $e_1.l > e_2.l$ and $e_2.offset < e_1.offset + e_1.vt <= e_2.offset + e_2.vt$ **then**
                add e_1 to $temp$;
            end if
        end for
    end for
    $curSet \leftarrow temp$;
    if $\frac{\sum_{n \in curSet} n.vt}{g_i.vt} < \beta$ **then**
        Anomaly is detected; break;
    else
        Put g_i at the end of SEQ;
    end if
end for
Extract the sub-trajectory that corresponds to SQE and assign it to tln;

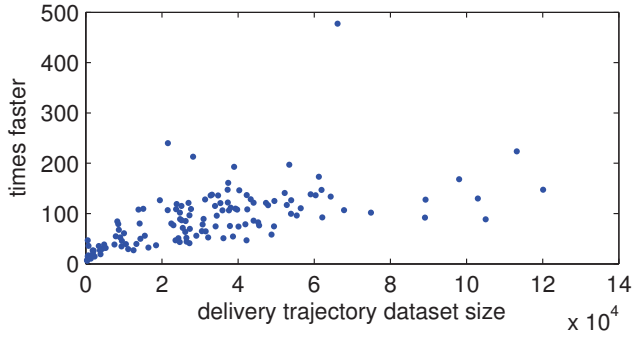


Figure 5.12: Improvement obtained with Inverted Index.

satisfy Equation 5.7 and 5.8. Then $|hasPath| = \sum_{n \in curSet} n.vt$ and $|passTraj| = g_i.vt$ and *Support* is calculated with Equation 5.4 for deciding the anomaly status. When the anomaly is detected, the longest normal grids are detected and the corresponding GPS samples are the longest normal sub-trajectory.

Efficiency Improvement We examine the efficiency improvement of the new algorithm with real life taxi data. The result is shown in Figure 5.12. We can see that, the new algorithm can achieves tens and even hundreds of times faster.

5.4.2.2 Passenger Finding Strategies after Dropping off Passengers

After detecting the initial intention sub-trajectories, we proceed to extract the passenger finding strategies of a driver after dropping off passengers. We use the same decomposition of the city as in Figure 5.9. For each time slot, we choose the top 99 busiest grids and the rest as one non-hot area and obtain 100 location labels. For each location label, we are interested in drivers' preference of three types of strategies after dropping off passengers, *i.e.*, *going distant*, *hunting locally* and *waiting locally*. The *local* and *distant* properties are judged by whether the distance between the first and last sample of the initial intention sub-trajectory d_{drop} is bigger than a threshold τ_d and the "hunting" and "waiting" are by whether the sub-trajectory is ended with waiting (the criteria is shown in Equation 5.9) event whose time duration t_{wait} is bigger than ω_d.

$$d_d \begin{cases} > \tau_d & \text{going distant} \\ <= \tau_d & \begin{cases} t_{wait} > \omega_d & \text{waiting locally} \\ t_{wait} <= \omega_d & \text{hunting locally} \end{cases} \end{cases} \tag{5.9}$$

For a specific location, we count the times of *going distant* event s_{dd}, *hunting locally* event s_{dh} and *waiting locally* event s_{dw}. In [60], we build the feature matrix directly with these numbers. However, as the total number of drop-off events of good drivers are significantly bigger than that of ordinary drivers, for most of the locations, the numbers

of committing specific strategies of good drivers are normally bigger than that of ordinary ones. And this difference can't reveal whether a driver prefers to this strategy. So in this thesis, we define a notion called strategy preference SP, which is measured by the percentage of events that adopt a strategy at a specific location l in time period t.

$$SP(s^{l,t}) = \frac{s^{l,t}}{s^{l,t}_{dd} + s^{l,t}_{dh} + s^{l,t}_{dw}},\ s^{l,t} = s^{l,t}_{dd},\ s^{l,t}_{dh}\ s^{l,t}_{dw} \tag{5.10}$$

. $SP(s^{l,t})$ indicates a driver's inclination to certain strategies, and it avoids the influence of different number of drop-offs among different drivers.

As each location has 3 strategies after dropping off passenger, for a driver, we build a 100 locations $\times$ 3 strategies = 300-dimension feature vector, in which each dimension corresponds to a specific $\langle location, strategy \rangle$ combination.

5.4.2.3 Passenger Finding Strategies before Picking up Passengers

Before picking up passengers, we measure whether a driver should wait or hunt around in one location. For a pick-up record, if the waiting time before picking up t_{wait} is bigger than a threshold ω_p, we say it is waiting, and otherwise it is hunting (Equation 5.11). We adopt the same formulation of location as passenger finding strategies after dropping off. And we build a 100 locations $\times$ 2 strategies = 200-dimension feature vector to enumerate the preference of each strategy.

$$t_{wait} \begin{cases} > \omega_p & hunting \\ <= \omega_p & waiting \end{cases} \tag{5.11}$$

5.4.2.4 Passenger Serving Areas

As shown in [66], high revenue drivers are capable of choosing to serve at certain city areas to make good profit and meanwhile avoid heavy traffics in Shenzhen, China. In this thesis, we also investigate the preference of passenger serving areas of different taxi drivers. Different from the passenger finding strategies, we decompose the city into 10×5 areas, with each one about $5km \times 5km$. For each driver, we count the passenger finding times pft_i in area i in each time slot. And the preference of area i is defined as:

$$P_i = \frac{pft_i}{\sum pft_i} \tag{5.12}$$

. which measures a driver's preference in serving in each area. For each time slot, we build 50-dimension feature vector, with each one the preference for a particular area.

5.4.2.5 Passenger Delivery Speeds

The passenger delivery speed is the average speed of all the passenger delivery trips in a particular time period of day. We calculate the passenger delivery speed $speed^h$ in hour h with the following equation:

$$speed^h = \frac{\sum d_{di}^h}{\sum t_{di}^h} \tag{5.13}$$

, in which d_{di}^h and t_{di}^h are the delivery distance and time of ith trip in hour h respectively. Even though $speed^h$ can't provide guidance to the drivers, a higher value of it implies good skills to choose the unobstructed route. And then for each time slot, we build a feature vector with each dimension for one hour.

5.4.2.6 Feature Document Preparation

For a time slot, we combine the aforementioned features in one row vector for each taxi and then accumulate the row vectors of different taxis into a feature matrix FM, in which each row corresponding to one taxi and each column for one specific strategy. The following analysis is based on this feature matrix.

5.4.3 Understanding Finding Serving Strategies

In this section we study the taxi serving strategies from the following three perspectives.

1. How can we learn from the strategies of good drivers?
2. What's the evolving trend of each strategy with the revenue? What strategies should be emphasized and what should be weakened?
3. What strategies can differentiate good and ordinary drivers?

Intuitively, the strategy inclinations of good drivers in certain time and location context are valuable experiences for other drivers to learn. So we first try to find the most inclined strategies of good drivers based on the feature document. Then we evaluate each individual strategy by measuring its correlation with drivers' profitability. The positive correlation value indicates that generally speaking, with more preference of this strategy, the drivers have higher profitablity, and vice versa. In the end, we investigate the strategical differences between good and ordinary drivers through classification method and try to find the strategies that can different the two driver groups.

5.4.3.1 Learning from Good Taxi Drivers

In a given location and time context, we want to learn the strategy that is mostly preferred by good drivers. To measure the preference of a particular strategy in a group of

drivers, we have to consider not only the preference of each driver, but also their profitability level, which is measured as the hourly revenue rate, and their experience, which is measured by the number of drop-off events in the corresponding location and time context. So we define a $Score(s_p^{l,t})$ function to measure the preference of a strategy $s_p^{l,t}$ of in a community of drivers.

$$Score(s_p^{l,t}) = \sum_{i=1}^{N} SP(s_i^{l,t}) * rev_i^t * s_i^{l,t} \tag{5.14}$$

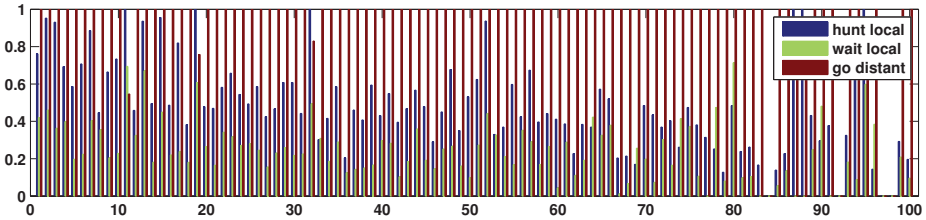
, which incorporates the three factors, *i.e.*, in given location l and time period t, for each good taxi i, $SP(s_i^{l,t})$ is its strategy preference of s, rev_i^t is its hourly revenue rate and $s_i^{l,t}$ is its number of drop off events at location l and time period t.

For the strategies after passenger drop-off, we calculate the *Score*s of going distant, hunting locally and waiting locally and convert them into range 0∼1 for each location and time context (shown in Figure 5.13). We choose the strategy with the maximum value as the most preferred strategy. Then for each time slot, we can draw a preferred strategy map which reveals the strategies learned from good drivers in each location (shown in Figure 5.14). Besides the detailed information, we can observe that, generally speaking, it's better to hunt locally in top hot areas and go distant in non-hot areas. And in most of the cases, hunting is better than waiting. However, there are some exception, such as the airport and tourist sites around the Western Lake in noon and afternoon time, where it's easy to wait for passengers at that time. The preference of waiting locally reduces when the hotness decreases. On the contrary, the preference of going distant increases in less hot areas as there are less taxi demands .

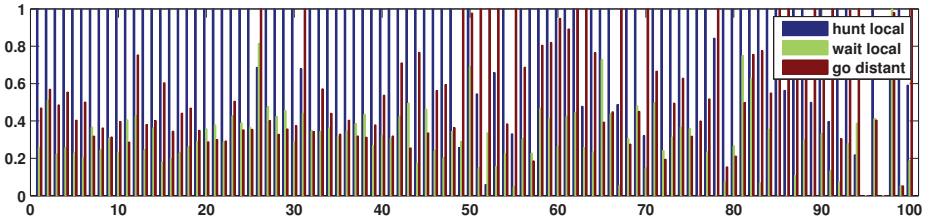
The learned strategies are the ones generally followed by good drivers in given location and time contexts. However, it's certain that if they are strictly followed by all the drivers, they will become bad ones because of the competition among the drivers. Actually in Figure 5.13 we can observe that except late night, when moving to less hot locations, the scores of going distant and local hunting become increasingly close. It indicates that, although hunting locally is the most preferred, going distant is not negligible as large number of drivers also prefer to it. In real life, how much degree a driver follows these strategies depends on their experiences and may be one key influencing factor of the revenue performance. So in next section, we intend to investigate how the degree of following one strategy influences the performance.

5.4.3.2 Evaluation of Individual Strategies

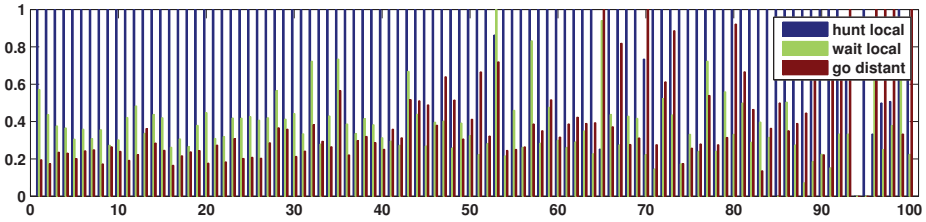
We evaluate each individual strategy by calculating the correlation value of the feature dimension with the corresponding drivers' profitability. A positive correlation value implies that, generally when drivers prefer more to this strategy, they earn more. So for the drivers



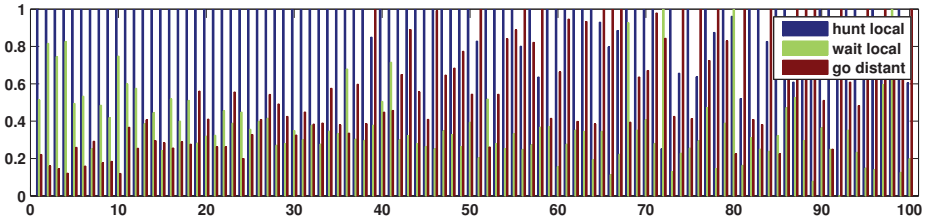
(a). 0:00∼5:59



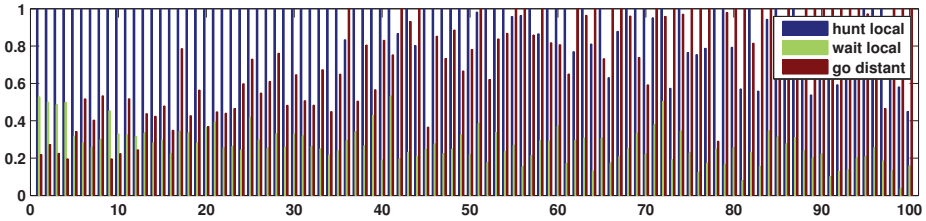
(b). 6:00∼9:59



(c). 10:00∼13:59

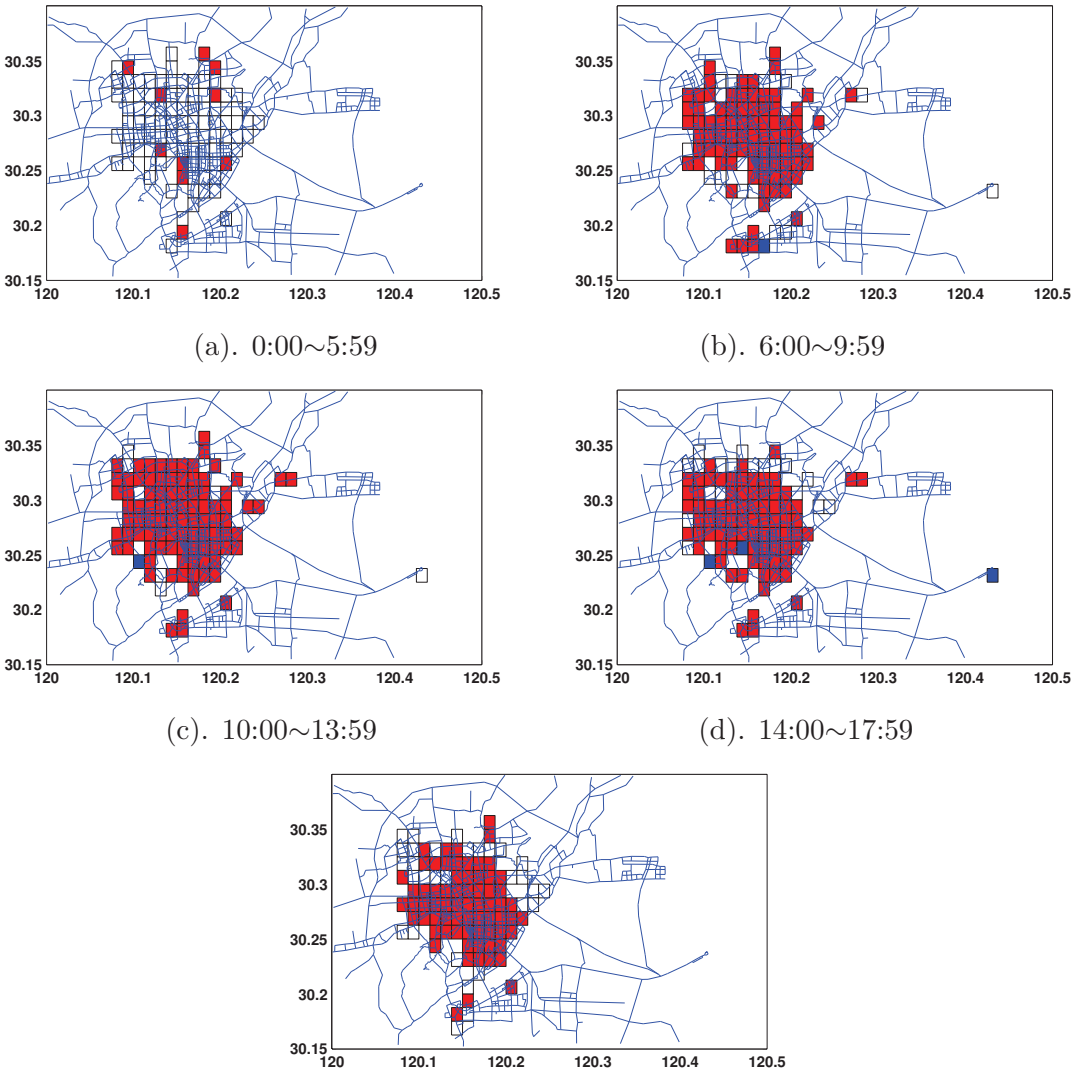


(d). 14:00∼17:59



(e). 18:00∼23:59

Figure 5.13: Strategy scores in different time slots.



(a). 0:00∼5:59

(b). 6:00∼9:59

(c). 10:00∼13:59

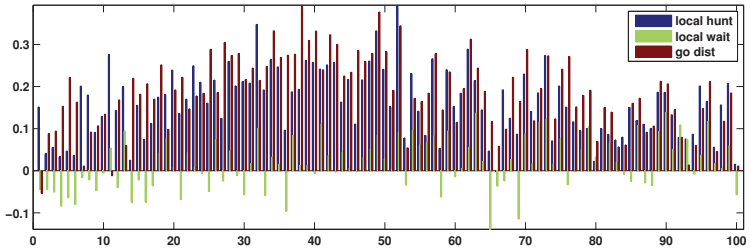
(d). 14:00∼17:59

(e). 18:00∼23:59

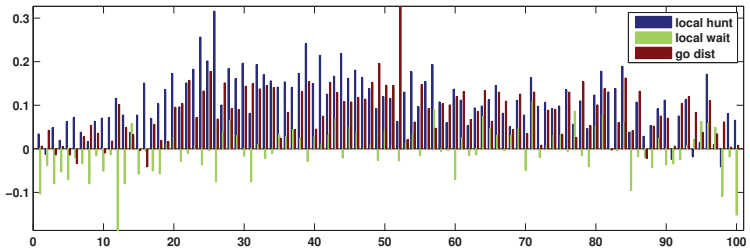
Figure 5.14: Strategy preference map. Red: hunting locally; Blue: waiting locally; White: going distant.

with lower preferences of the strategy, they'd better increase it. On the contrary, with small negative correlation value, the drivers with bigger preference should decrease it. For each feature dimension f and the hourly revenue r of corresponding drivers, we calculate the correlation $corr(f, r)$ with Equation 5.1. The result is revealed in Figure 5.15, 5.16 and 5.17.

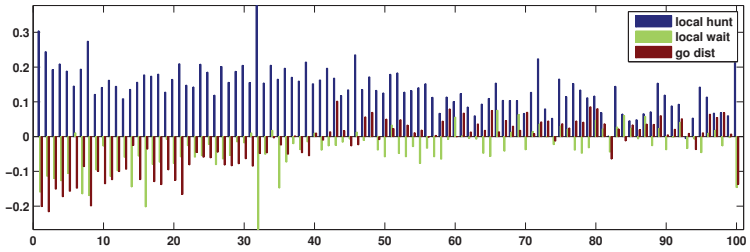
For the passenger finding strategies after drop-off, one can easily observe that local waiting normally has negative correlation value for the hot areas, indicating that to make



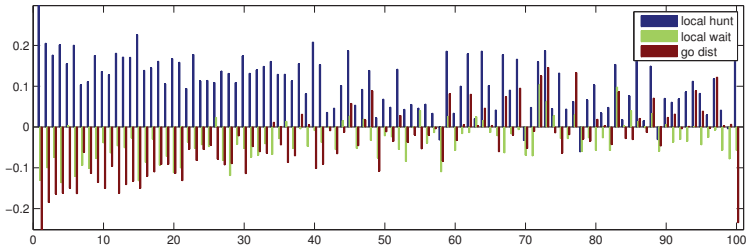
(a). 0:00∼5:59



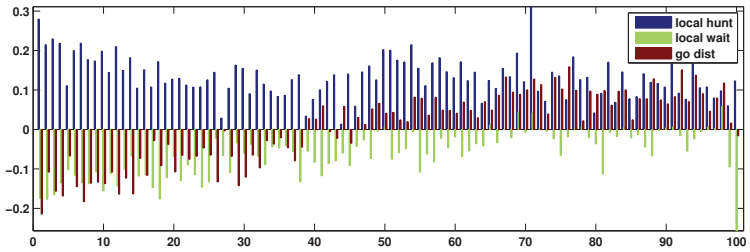
(b). 6:00∼9:59



(c). 10:00∼13:59



(d). 14:00∼17:59



(e). 18:00∼23:59

Figure 5.15: Correlation of passenger finding strategies after drop-off and the revenue in different time slots. Horizontal axis is the top 1∼99 hot drop-off locations and the rest area (labelled as 100).

more profits, taxi drivers should prefer less to local waiting there. On the contrary, for most of the location and time contexts, local hunting is positively correlated with the revenue, which means generally more local hunting causes more revenues. This observation complies with the experiences learned from good drivers that local hunting is better in most of the cases by learning from good drivers. And it is also consistent with the results of the strategies before pick-up (Figure 5.16). For the majority of the locations, hunting before pick-up is positively correlated with the revenue while waiting is negatively. Meaning in these areas, taxi drivers should prefer more to hunting and decrease waiting events. From noon until night, going distant is negatively correlated with the performance in the top hot areas and the value moves to positive when moving to less hot areas. It also complies with the previous results that taxi drivers should always focus on hot areas.

The correlations between taxi serving areas and the revenues are shown in Figure 5.17. The red area in Figure 5.17 covers night entertainment areas (night clubs and etc.) and the residential areas. Taxis serving more here obtain more revenues generally. The dark blue areas are where taxis generally earn less if they serve more and are mostly the suburb areas.

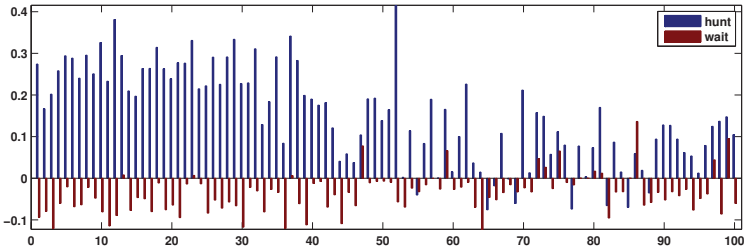
One thing that should be pointed out is, the absolute correlation value is not close to 1, which is because of the fact that revenues are influenced by a variety of practical factors together with the serving area, such as the traffic, driver preference, passenger distribution and variance. But still it can provide indications of the profitability trend when the drivers serve more or less in each area and be used to guide taxi drivers.

5.4.3.3 Differentiating Good and Ordinary Taxi Drivers Based on the Strategies

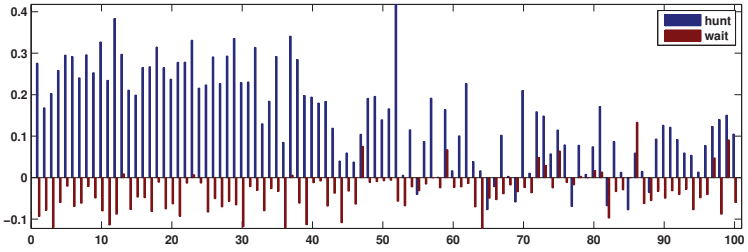
The aim of this section is to find the group of strategies which can differentiate good and ordinary taxi drivers, so that we can have a clue of what differentiates them. We choose the top 500 profitable drivers as the good drivers and the middle 500 taxi drivers as the ordinary drivers. We label the features vectors of good and ordinary drivers with 1 and -1 respectively, and use *L1-SVM* [9] and *AdaBoost* [103] to classify the two groups. The weaker classifiers in *AdaBoost* are binary classifiers in each feature dimension:

$$Label(f_i) = \begin{cases} 1, & f_i \geq \gamma \\ -1, & f_i < \gamma \end{cases} \tag{5.15}$$

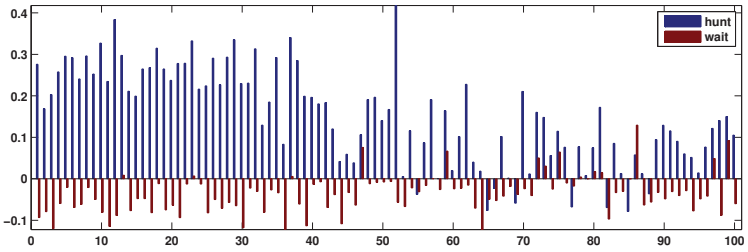
. For each time slot, we randomly divide the drivers into 5 equal-sized groups and use 5-folder cross validation to get the classification accuracy. The results of *L1-SVM* and *AdaBoost* with different number of weak classifiers in different time slots are shown in Table 5.7 and Figure 5.18 respectively. We can see that *L1-SVM* achieves about 85∼90% accuracy



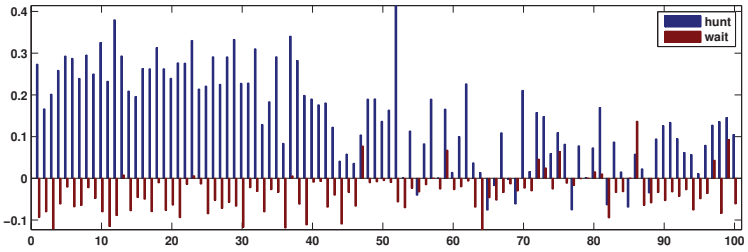
(a). 0:00∼5:59



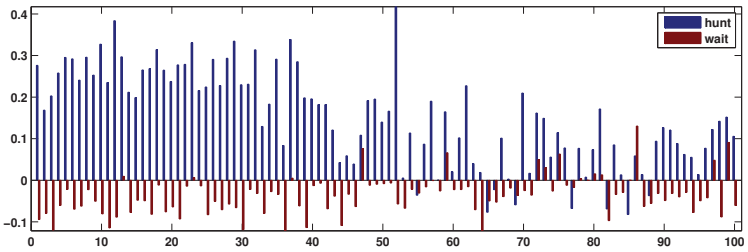
(b). 6:00∼9:59



(c). 10:00∼13:59

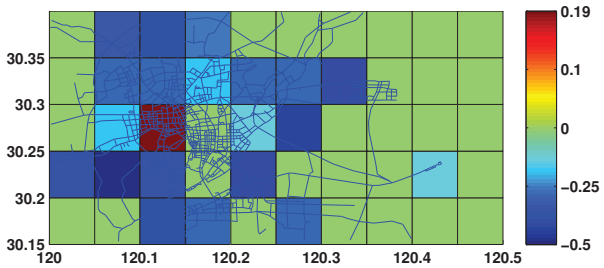


(d). 14:00∼17:59

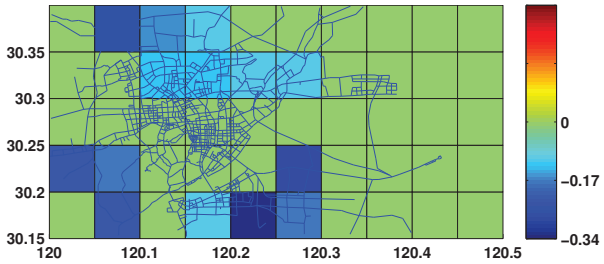


(e). 18:00∼23:59

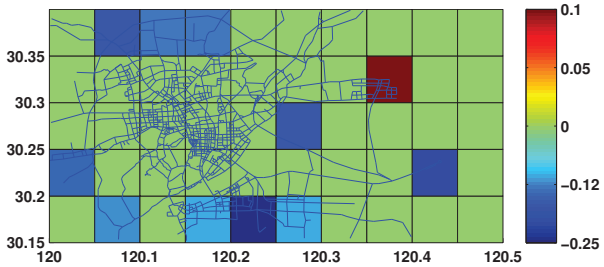
Figure 5.16: Correlation of passenger finding strategies before pick-up and the revenue in different time slots. Horizontal axis is the top 1∼99 hot drop-off locations and the rest area (labelled as 100).



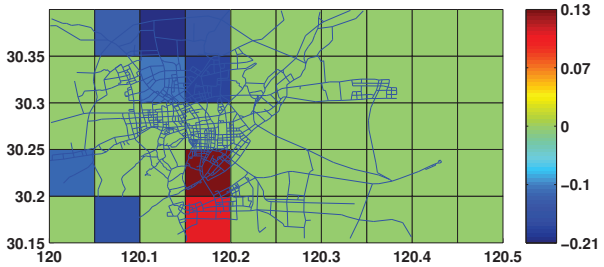
(a). 0:00∼5:59



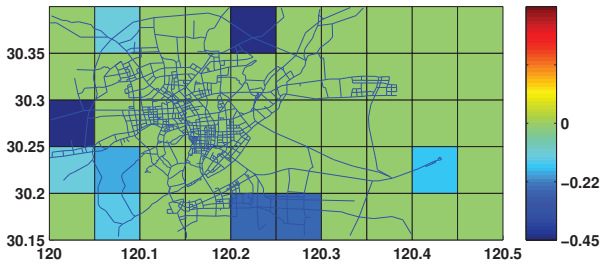
(b). 6:00∼9:59



(c). 10:00∼13:59



(d). 14:00∼17:59



(e). 18:00∼23:59

Figure 5.17: Correlation of taxi serving areas and the revenue in different time slots.

Table 5.7: Classification accuracy (%) of L1-SVM

night	morning	noon	afternoon	evening
89.3±1.5	89.7±2.8	89.2±2.5	87.3±2	85.8±1.7

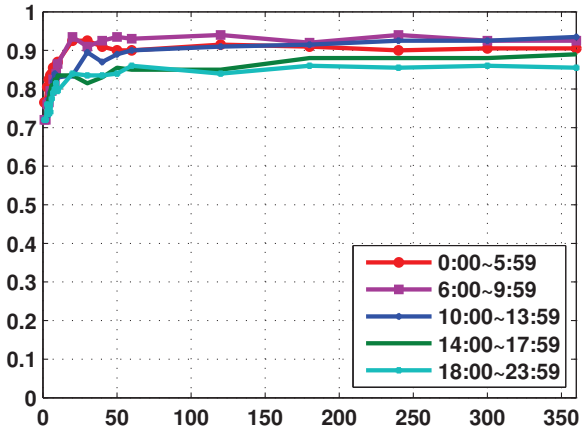


Figure 5.18: Classification accuracy of AdaBoost vs. different number of weak classifiers.

and *AdaBoost* achieves about 86∼94%, implying that there do exist some strategies that can differentiate good and ordinary drivers.

With *L1-SVM*, we are able to obtain a set of salient features which can differentiate the two driver groups. And their contributions are measured by the assigned weights during the classifier training phrase. These salient features are believed to be useful for suggesting efficient passenger-finding strategies in [60]. In this thesis, we further study these strategies individually by examining their mutual information [34] with the revenue.

Much work used information theoretic criteria in feature selection [7, 23]. The mutual information between a feature f and the target y is defined as:

$$I(f) = \int_{f_i} \int_{y} P(f_i, y) \log \frac{P(f_i, y)}{P(f_i)p(y)} df dy \tag{5.16}$$

, in which $p(f_i)$ and $p(y)$ are the probability densities of f_i and y. Bigger information value means better differentiating power of f over different ys. In this thesis, ys are the two labels and f_i is preference of strategy f of taxi i. We discrete the range of f into 20 equal-size intervals and Equation 5.16 becomes to be:

$$I(f) = \sum_{f_i} \sum_{y} P(F = f_i, Y = y) \log \frac{P(F = f_i, Y = y)}{P(F = f_i)P(Y = y)} \tag{5.17}$$

, where $P(i)$ can easily be calculated by counting the occurrences.

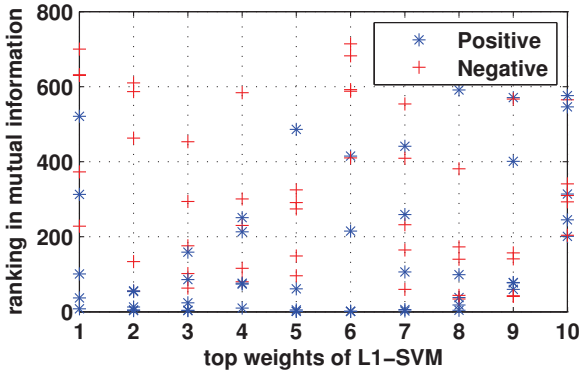


Figure 5.19: Rankings of the top 10 positive and negative features in *L1-SVM vs.* rankings of their mutual information in all time periods.

The mutual information rankings of the top 10 positive and negative features in *L1-SVM* in all time periods are shown in Figure 5.19. We can see that, the information rankings of the majority positive 10 features are within top 50, indicating that the majority of these features possess high differentiating power. However, there are several features with high *L1-SVM* weights but low information ranking. This is because of that: "a variable (*i.e.* feature in this thesis) that is completely useless by itself can provide a significant performance improvement when taken with others" [34]. And all these features together reach good differentiating results between the good and ordinary taxis instead of simply adding their differentiating power together.

With both *L1-SVM* and *AdaBoost*, we obtain more than 100 features with non-zero weights in each time slot, based on which we classify the good and ordinary taxis. It reveals that good drivers achieve better performance not by a single strategy, but by a combination of a group ones. By accumulating the advantages through all of them, good drivers achieve outstanding performance compared with ordinary ones.

5.5 Conclusion and Discussion

The digital footprints of a community of drivers are valuable resources for us to study the characteristics of taxi drivers' behaviors and thus to understand the hidden human intelligence under different context situations. In this chapter, we provide a thorough analysis of a large number of anomalous passenger delivery behaviors and work shifting behaviors collected from a community of taxis. we perform analysis aiming to answer the following questions. a) What percentage of all trips are anomalous? b) Out of the anomalous trajectories, what percentage of them travel longer distance than necessary? c)

What statistical "tendencies" can we discern from the detected anomalous trajectories? d) Do taxi drivers who have a higher tendency to commit fraud have an economical advantage over those who don't? We observe that 1) Over 60% of the anomalous trajectories are "detours" that travel longer distances and time than normal trajectories; 2) The average trip length of drivers with high-detour tendency is 20% longer than that of normal drivers; 3) The length of anomalous sub-trajectories is usually less than a third of the entire trip, and they tend to begin in the first two thirds of the journey; 4) Although longer distance results in a greater taxi fare, a higher tendency to take anomalous detours does not result in higher monthly revenue; 5) Taxis with a higher income usually spend less time finding new passengers and deliver them in faster speed. Besides, we analysis the spatial and temporal distributions of all the work shifting events. We find that, afternoon work shiftings normally happen between 16:40∼17:20 in non-hot areas, which explains that why people feel hard to get taxis in this time period.

We also study the taxi serving strategies based on a real life large scale taxi dataset collected from 7600 taxi in China, aiming to discover both efficient and inefficient techniques, and reveal the underneath facts. To learn the initial intention of drivers right after they drop off passengers, we propose a novel method to extract the initial intention sub-trajectory by comparing with the passenger delivery trajectories and improve the efficiency with tree compression and inverted index searching. Then we formulate the strategies concerning passenger finding behaviors after dropping off passenger and before picking up passengers, passenger serving areas, and delivery speeds. By learning from the top performance taxi drivers, we provide a strategy map, revealing the good strategies at each place in different time periods of day. Through calculating the correlation relationship between each strategy and the revenue, we measure their influences on performance. Generally we observe that hunting is better than waiting locally in hot areas with some exceptions including the airport and some tourist sights. Despite the influence of heavy traffic in hot areas, going distant is not a good choice. When moving to non-hot areas, they should go to hot areas (going distant). We further classify the good and ordinary drivers based on the proposed features,

Chapter 6

A Smart Taxi System

Contents

6.1 Smart Taxi System Introduction

Enriched by the perception and understanding of human behaviors in scales of individual and community, the service scope and functionalities of the smart sensing systems will be greatly expanded. In this chapter, we illustrate this trend with a smart taxi system, which targets at not only dispatch and navigation services for drivers and passengers, but a complete and considerate taxi service system that allows passenger to easily and comfortably enjoy taxi services and assist taxi drivers to in their business process. Besides, it also adopts the fruits born from the exploration of the digital footprints, especially the monitoring and understanding of human behaviors, and supports various types of value-added services to other possible clients, such as taxi companies, city traffic administrative bureaus and the public.

Currently, GPS-based taxi dispatch systems are widely adopted in big cities in China for location-based taxi dispatch service which can greatly improve the taxi service efficiency. With real time collections of the taxi locations and status, it's easy for taxi dispatch center to monitor the locations of the vacant taxis. With the hotlines for people to call for taxi services, the dispatch center collects the passenger demands and dispatches nearby taxis to satisfy them. With the advancement of GPS-embedded smart mobile phones, we can design easy-to-use mobile application which automatically gathers the location of the passenger and negotiate with the dispatch center to get a delivery efficiently.

Besides taxi dispatches, the smart taxi system actually can work as agents for both taxi drivers and passengers to assist the whole taxi service process. For taxi drivers, it can recommend efficient passenger finding suggestions, respond to nearby passenger calls, navigate passenger deliveries and even automatically handle the payment process. While for passengers, it can provide "one-click" service which automatically gathers the passenger location and talks with dispatch center to get a taxi ride. Besides, it also can guide passengers to locations where it's easier to get taxis in case of rush hours. And by monitoring driver's anomalous behaviors, it's able to provide anomaly alerts if required. To save the cash payment trouble, it can maintain money accounts for both taxi drivers and passengers. After the delivery is completed, it automatically transfers the payment from the passenger's account to the driver's to save the trouble of money delivery.

By exploring the GPS footprints obtained in the system, the smart taxi system also provides smart services to taxi companies, city traffic administrative bureaus and even the public. For example, with the anomalous passenger delivery detection service, taxi companies can monitor all the anomalous behaviors of taxi drivers and take necessary measures to reduce such behaviors and improve the taxi service. The hotspots and human mobility patterns revealed from the GPS traces can serve as data sources for city planners to design public transportations and for city governors to monitor the abnormal changes in the city. As taxis are traveling all over the city, whose speeds reflect the traffic situation on the road, their real time reports provide a perfect picture about the city traffic situations. Besides, the novel path planning solutions (such as *T-Drive*) based on large number of passenger delivery paths provide better navigation services to the public than conventional methods by incorporating the human intelligence of the experienced drivers.

The working scenarios of the system is illustrated as in Figure 6.1. It consists of four roles. The first one is the smart taxis, which adopt mobile phones or other professional taxi dispatch equipments to send GPS reports to a central server and assist the taxi service process. The second one is the smart passengers, which utilize mobile phone clients to communicate with the central server for taxi delivery services. The third one is the smart monitors, which includes other possible clients, such as the taxi companies, traffic bureaus

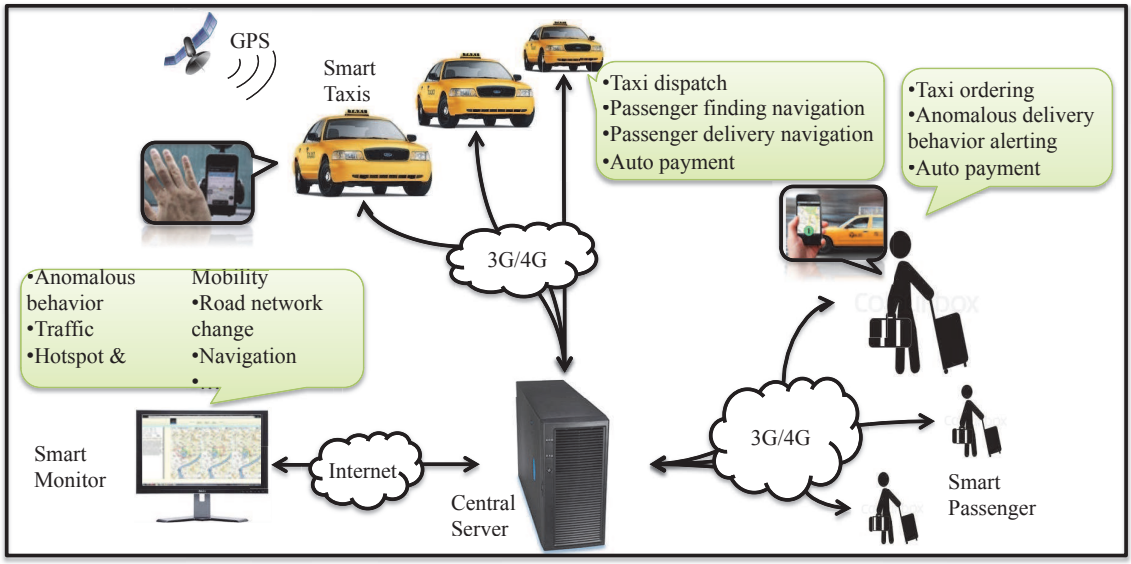


Figure 6.1: Working scenarios of the smart taxi system.

and navigation service companies. Based on the detected human behaviors, they can monitor the taxi service, traffic conditions, city hotspots, human mobilities, and so on. The last one is the central server, which is the core of the system. It gathers the GPS reports from taxis and takes in charge of the services to the other three roles.

The conceptual framework of the system is shown in Figure 6.2, which shows the inner functions and service sequences. Here we introduce the roles and functions of each parts in detail.

Smart Taxi: The smart taxi client provides assistance to taxi drivers during their business, such as dispatching taxi drivers to nearby passenger calls and assisting them in the passenger finding and delivering process. Firstly, during the passenger finding process, it provides suggestions to the drivers to efficiently find passengers in current contexts (location, time and weather) by incorporating hotspots, mobility patterns, or/and detailed guidance like in [123], or/and strategical guidance like what has been presented in this paper. It also displays the real time delivery demands issued by the passengers in the Smart Passenger client to the drivers. After picking up passengers, it provides efficient ways to deliver the passenger to their destinations. Meanwhile, after dropping off passengers, it provides smart payment service, of which the system calculates the traveling fee and automatically charges moneys from the passenger account to the driver account. It saves the money delivery process from both sides.

Smart Passenger: The smart passenger client provides considerate taxi services to passengers. With it, passengers only need to press a button to issue a taxi delivery demand.

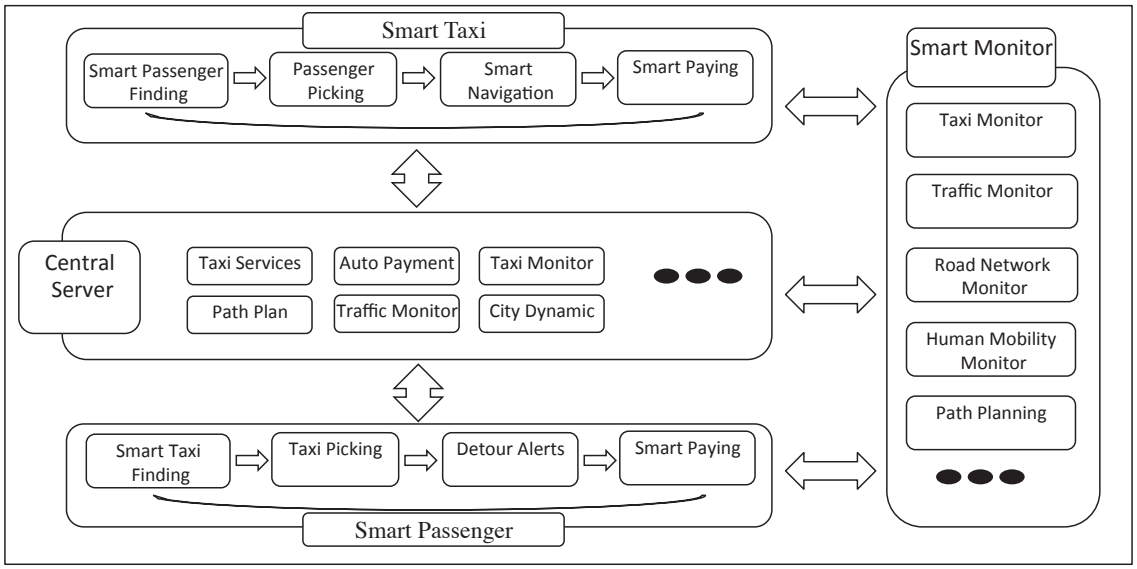


Figure 6.2: Framework of the smart taxi system.

It automatically collects the GPS location of the user and sends it to the central server within a taxi delivery request. For simplicity, the server can automatically dispatch the closest vacant taxi to serve. Meanwhile, we also can adopt the following easy way to select the taxi with better service quality. When receiving a delivery request, the server automatically replies with a few nearby vacant taxis with their service rankings obtained by the evaluations of previous passengers and the estimation of traveling time to the passengers' location. Passengers with the smart passenger client can choose the one she/he wants and negotiate directly with him to get a taxi ride. Before picked up, the smart passenger client may show the real time location of the taxi to reduce possible waiting anxiety. During the delivery process, in case the driver is not following a normal route, the client is able to issue an alert to the passenger. After dropping-off, it automatically pays the taxi service. But it needs a prepaid account of the passenger with sufficient amount of money in it.

Smart Monitor: The smart monitor client includes all the services that the smart taxi system provides to clients other than taxis and passengers. In Figure 6.2, we list some of the possible applications, including the taxi monitor, traffic monitor, road network monitor, human mobility monitor and path planning service. Taxi monitor provides the real time locations, passenger status and the moving speed of all taxis to taxi companies and city traffic bureaus. It also monitors the anomalous drivers in real time and provides possibilities for taxi companies to take measures to reduce fraudulent driving. Besides, it's also useful for city traffic bureaus to monitor unexpected events like taxi strikes. The traffic monitor service displays the traffic status of the road network derived from on-line taxi GPS

packets and the anomalous traffic patterns which might indicate traffic accidents. The road network monitor refers to the ability to detect the road network changes, including the opening (*e.g.*, newly-built) or blocking of road segments or even building the road network from scratches. Human mobility monitor reveals the hotspots and the real time human mobility among different areas in a city. It's able to support passenger finding and public transportation design. Path plan service provides efficient traverse routes to the public, which are mined from the large-scale passenger delivery behaviors of taxis.

Central Server: The central server is the core component that detects the human behaviors and provides services to the other three types of clients. The main data source of it is the GPS reports received in real time from the taxi fleet. It maintains the evaluations generated by previous passengers for each driver. When receiving a delivery demand initiated by a passenger client, it offers a few of the vacant taxis nearby with their evaluation rankings and handles the following communication between the passenger and the taxi driver. Meanwhile, it offers delivery navigation service, performs the real time anomaly delivery behavior detection and provides alerts to drivers if is required by the passenger. It also maintains an account for each driver and passenger, so that it can automatically pay the driver from the passenger's account, saving their effort for the cash delivery. After drop-off event, it provides the passenger finding guidance to taxi drivers.

Another main function of the central server is to detect the human behaviors, perform analysis and provide services to the smart monitors. It can either handle the data process and deliver the results directly to the smart monitors, or it can acts as data feeds of another smart system that handles the services of smart monitors. For example, the central server can collect the hotspots directly and visualize it to the city traffic administrative bureau. It also can collect the human mobility pattern and provide the data to the city planners who based on such information design publication transportations.

The challenges of such a system are mainly brought by the huge number of taxis and passengers, as it needs to monitor thousands of taxis in real time. So the system has to be carefully designed to meet the real time responses. And with the development of the city, new taxis will enter the market and thus the system should be salable. We envision this type of system can be deployed in larger cities, such as Beijing (around 70,000 taxis) and Mexico city (around 250,000 taxis), so it is necessary to verify the scalability of the proposed system.

6.2 Architecture of Smart Taxi System

We mainly introduce the design of the central server in this thesis, as it handles all the complicated data processing work such as the activity recognition and analysis tasks. The

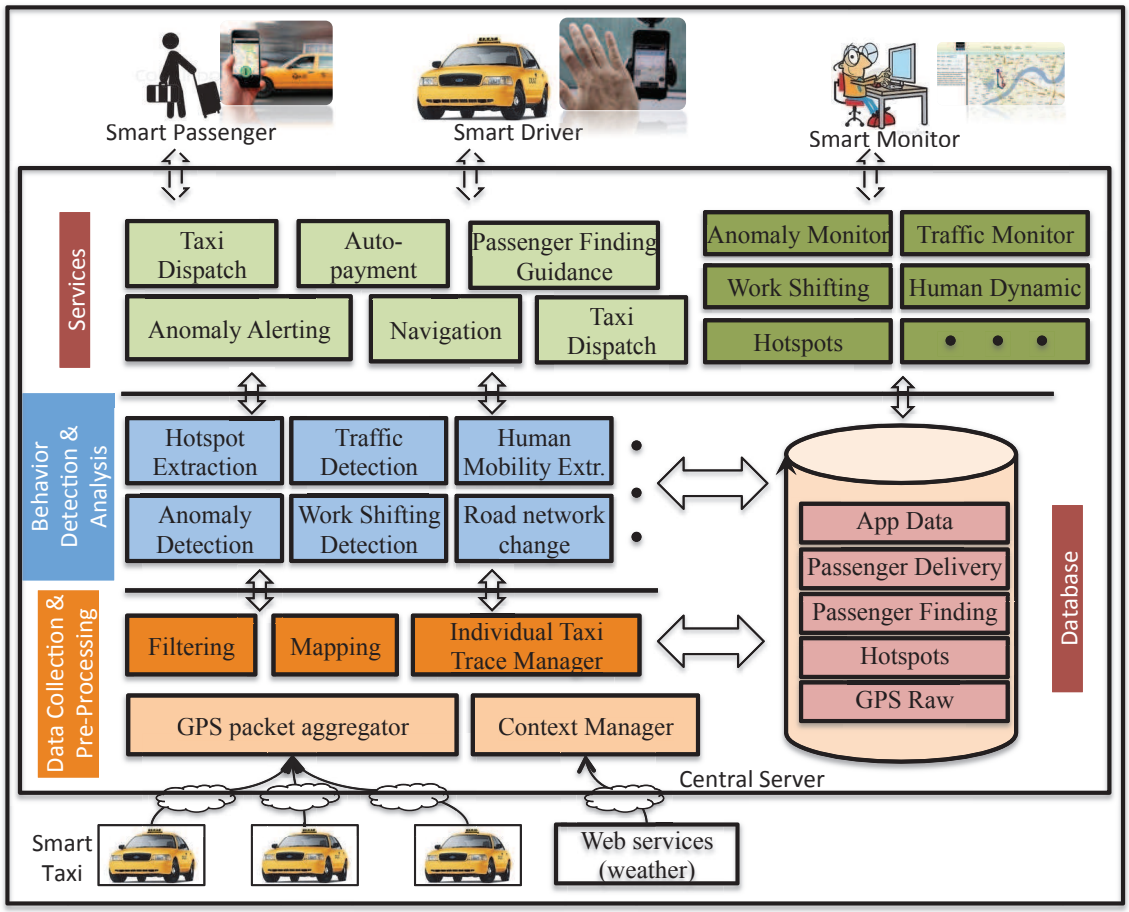


Figure 6.3: System architecture of the smart taxi system.

architecture design of it is shown in Figure 6.3, which includes mainly four modules, *Data Collection and Pre-Processing*, *Behavior Extraction*, *Database* and *Service*. *Data Collection and Pre-Processing* module receives the GPS packets from all taxis, tracks the operations of each individual taxi and performs data filtering and mapping tasks. It also collects the context information from network services (such as the weather service of Yahoo), which are necessary for applications like anomalous passenger delivery detection. The collected data is store in the *Database* module, which stores and manages the raw GPS records for all taxis and the extracted hotspots, passenger finding and delivery trajectories as well as the application specific data. The *Behavior Extraction* module extracts the individual and community behaviors from the collected records. The results of this module can be used in the *Service* module, which provides services to the three clients. We elaboration the details of these modules in the following sections.

6.2.1 Data Collection & Pre-Processing

6.2.1.1 GPS Packet Aggregator

GPS Packet Aggregator provides an interface that receives the GPS reports sent from the smart taxi clients in taxis, whose format is shown in Table 3.2. To effectively gather data, we adopt the Representational State Transfer (REST) architecture, which represents each resource (in our case taxis) as an URL. The clients of our system use HTTP POST operation to upload the GPS packet information into our system with this URL. For example . After receiving the data, the GPS Packet Aggregator extracts the taxiID from the URL and other parameters from the POST message body to form a complete record.

6.2.1.2 Context Manager

The context manager mainly takes care of the environmental factors that influence the human behaviors and currently it just manages the weather, but it can easily include other factors if necessary. The reason why we don't use the weather information directly is because practically we don't deal with the weather information directly but use some classification of it to obtain more general abstractions, such as "server" and "normal". For example, in the anomalous passenger delivery detection function, we don't need to consider the weather conditions exquisitely as their influences are not so subtle. Each context type is assigned with a unique ID. When queried, it can reply this ID to the requester.

6.2.1.3 Data Filtering Component

Data filtering component fulfills the data filtering tasks that are needed to handle the problems mentioned in Section 3.3.3, mainly including the "data entry erroneous" and "improper occupied flag". To handle the unreasonable GPS jumps, we calculate the average traveling speed between the current GPS packet and the previous one. If the speed is higher than a threshold, it means we need to filter out it. To handle the improper occupied flag problem, we setup three filtering criteria. The first one is whether the time duration of a passenger delivery trip is lower than a minimum or longer than a maximum threshold. The second one is whether the traveling distance of a trip is too short (e.g., less than $200m$, very suspicious for real life deliveries). And the last one is whether the vacant time duration while the taxi is moving is longer than a threshold. If the trip satisfies any of these criteria, we label it as suspicious and leave the processing logic to the applications. If the percentage of "suspicious" time duration in a day is too high for a taxi, we will label it as a "suspicious" taxi.

6.2.1.4 Mapping & Augmentation

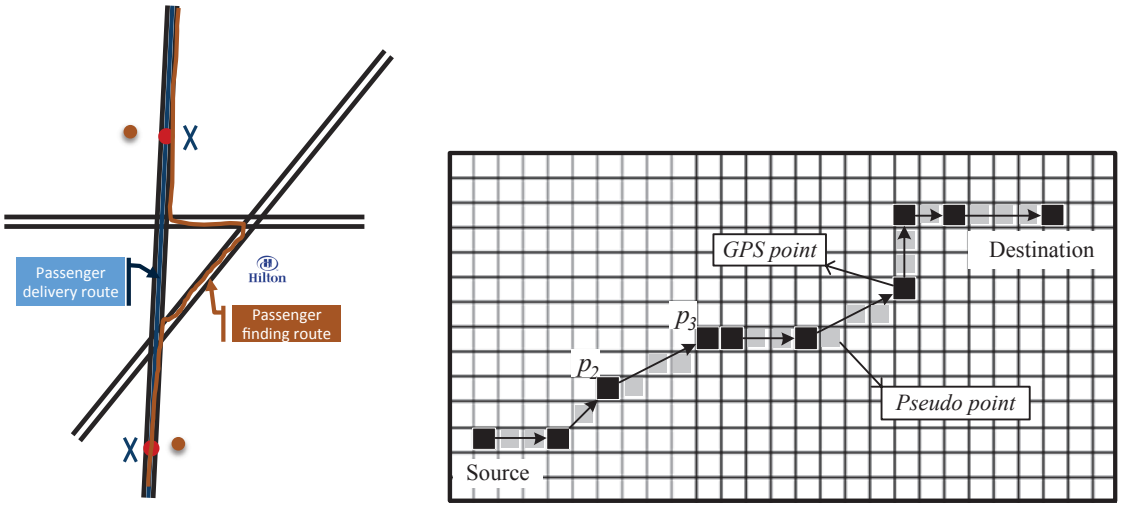
As we mentioned before, in order to feasibly manage the GPS points, we construct a *finite* abstract representation, or decomposition, of the original two-dimensional GPS plane. The Mapping component maps a GPS point into one of the element of the chosen decomposition where it "lands' in. Additionally, the Mapping component can also map a completed trajectory into a sequence of the decomposition elements. However, due to the low-sampling-rate problem, the successive elements may not be adjacent. The purpose of augmentation is to ensure that there are no gaps between successive mapped GPS points in a trajectory.

There are a number of ways to decompose the city, such as the grid decomposition, digital road network mapping, or partitioning it into different functional regions. In road network mapping, we map the GPS samples into the road segments where it most likely traverses. As is shown in Figure 6.4(a), the two cross points are the GPS sampling locations in a passenger finding trajectory. They are mapped as the red dots in the road segments. In grid decomposition, we split the city area into a matrix of grid cells (the grids shown in Figure 6.4(b)), and each GPS point maps to the grid cell where it lands (the black squares in in Figure 6.4).

Normally the augmentation is only performed for the passenger delivery trajectories, as they usually follow efficient routes for quick deliveries. For road network mapping, we can compute the shortest route between the adjacent mapped points (such as the passenger delivery route shown in Figure 6.4(a)). While for grid decomposition, we can insert pseudo cells along the line defined by the adjacent two grids. For example, in Figure 6.4(b), three pseudo cells (shown as gray cells) are inserted between p_2 and p_3. Eventually, such an augmenting process will allow us to obtain a cascaded cell sequence for the representation of each trajectory. However, such augmentation methods aren't suitable for passenger finding trajectories, whose primary purpose is for finding passengers. For example, the two brown dots in Figure 6.4(a) are the sampling locations in a passenger finding trajectory. Instead of following the shortest path, it takes the brown route because there is a Hilton hotel, where it's easy to find passengers.

6.2.1.5 Individual Taxi Trace Manager

Individual Taxi Trace Manager tracks the GPS footprints of each taxi, extracting passenger finding trajectories, passenger delivery trajectories and the pick-up and drop-off locations. The obtained results are saved in the database that can be directly used for the other modules. In practice, we manage a GPS packet queue for each taxi. If the queue is empty or the state of the new arrived packet is the same as the ones managed in the queue,



(a) (b)

Figure 6.4: Illustrative examples of mapping and augmenting a trajectory. (a) road mapping; (b)grid decomposition.

we just put the packets in the queue. Otherwise, it implies a passenger pick-up (from vacant to occupied) or drop-off (from occupied to vacant) event. The trajectory managed in the queue is a passenger finding or passenger delivery trajectory respectively. It is saved in the database and the queue is cleared after.

6.2.2 Database and DB Manager

The database stores and manages all the raw GPS records for all taxis. Besides, it also keeps the extracted hotspots, passenger finding and delivery trajectories, all of which have been labeled with contextual information provided by the Context Manager, so that the applications can choose the records with appropriate type of context easily. For the fast retrieval of passenger finding trajectories, which are widely used for the anomalous passenger delivery behavior detection and the initial intention detection, we make use of the inverted index mechanism [132] as well. We maintain a secondary database whose records are the mapped grid cells and the elements of each record are trajectory-position pairs, indicating the trajectories where the indexing grid cell appears, along with its position in that trajectory. Then to select the trajectories traversed from one cell to another, we compare the pairs of the two cells to choose the trajectories that have records in both grids in the correct order. The operations are conducted in the DB Manager.

6.2.3 Behavior Extraction

The Behavior Extraction module is in charge of detecting and analyzing the individual and community behaviors from the collected GPS traces. We introduce the main components presented in the system architecture in the following sections.

6.2.3.1 Hotspot Extraction

The hotspots are the distributions of the pick-up and drop-off behaviors in a taxi fleet. Hotspots Extraction module takes in charge of extracting the hotspots from the large collected pick-up and drop-off events. One easy way to do it is by simply counting the number of such events in each cells in the grid decomposition. Some typical results are shown in Figure 5.9. The other ways are using the clustering methods, such as the *K-means* [73], *Agglomerative Hierarchical Clustering* [78] and *DBSCAN* [93] to cluster the events into a number of clusters.

6.2.3.2 Human Mobility Extraction

With the obtained large number of passenger delivery trips, human mobility model measures the linkage strength between two different areas. Practically, given the coverage of two areas, the linkage strength can be simply defined as the number of passenger deliveries from one area to the other in certain given context, such as time of day and day of week. So the human mobility pattern depends on how we choose the two areas to measure. Normally we measure the human mobility among the hotspots of the city. We count the number of passenger deliveries among the hotspots and rank them in descending order. Then we can obtain the top strong human mobility patterns.

6.2.3.3 Traffic Detection

When the taxis deliver passengers, they can be considered as running sensors that perceive the city traffic conditions. By the real time aggregation of the speed information of a large taxi fleet, we can map the records into the road segments and obtain a comprehensive picture about the real time city traffic. For a road segment with several taxis, we can first exclude parked taxis (they should stop for a relatively long time) and then average the speeds of the rest taxis.

6.2.3.4 Anomalous Passenger Delivery Detection

To detect the anomalous passenger delivery behaviors in real time, we must know the destination information at the start of the trip. The GPS locations of the taxis have to be reported to the server if they want to use the smart dispatch service, and this benefit

could compensate the driver's unwillingness to be checked with their anomalous behaviors. To employ the anomalous trajectory detection, the system needs to know the destination before the trip, so that it can judge whether the trip is anomalous or not. There are many possible ways to obtain this information and we just offer one simple way, pressing the destination region for over three seconds. Other opportunities include the chance when they require the navigation services, or the passengers explicitly require the services. With the advancement of voice recognition, we are also expecting feasible vocal ways to input by audio. It's worth noting that, the destination area needn't to be precise, because *iBOAT* only needs the destination area so that it can gather the historical trajectories that share the same source and destination areas.

The anomalous passenger delivery detection module aims at detecting the anomalous passenger deliveries for a large number of taxis simultaneously in *real time*, the primary issue we have to address is the response time of the system. In a city like Hangzhou, China (where our dataset is taken from), there are 8,500 taxis operating in the city. Thus, our system should be able to perform anomaly detection on all 8,500 occupied taxis *simultaneously*, at a rate of at least once per minute (taxi GPS sampling time).

In the process of anomalous trajectory detection, another critical issue is to consider the effect of context such as time and weather during the preparation of historic trajectory set. Since our system is based on historical "traces", it is important to consider only those historical trajectories that occurred under similar context. It should be noted that previous anomaly detection mechanisms have only considered physical location as a relevant context for accumulating historical trajectories. We believe that additional contextual information will improve the overall accuracy of the system.

As elaborated in Section 4.3, we adopt the inverted index mechanism and greatly improve the recognition efficiency of the *iBOAT* method. Here we explain how to adopt this method to conduct the anomaly detection in real life. As passengers may travel among different areas, there are large number of possible $\langle S, D \rangle$ pairs. In practice, it will be much time costly to build the historical trajectory dataset every time for each passenger delivery trip. So we build the dataset for all the possible $\langle S, D \rangle$ pairs and store them in the database. Instead of dealing with the trajectory dataset, we directly store the inverted index dataset of it and update it at a low frequency like once per day. When testing a new trajectory, we query the corresponding inverted index dataset directly from the database and perform the anomalous behavior detection task.

6.2.3.5 Work Shifting Detection

Work Shifting Detection module is in charge of detecting the work shifting events of a taxi. For each taxi, we manage work shifting agreement (the agreed location and time

period, the detection method is introduced in Section 4.4) in the database. When the time period is approaching, we monitor the event that the taxi goes to the work shifting location and stays a while and consider as the event as the work shifting. After the work shifting event, the owner of the digital traces is changed.

6.2.4 Services

The service layer includes the components that take in charge of the processing of different applications based on the raw sensing records or the extracted individual and community behaviors. As these services focus on designing easy-to-use applications (interfaces and so on) that cater for the real life demands, they are not the main research target of this thesis. Here we present some easy design examples of the applications to illustrate the simple use scenarios. To make the introduction more understandable, we also present the possible interfaces or logics for the three clients.

6.2.4.1 Taxi Dispatch

Taxi dispatch is the initial purpose of deploying GPS devices in taxi, and it has been commercialized for a few years in many cities. Here we present a simple application that takes advantage of people's mobile phones to provide location-based dispatch services. It works as follows in our prototype. When a passenger needs to take a taxi, he launches the smart taxi application in his mobile phone and login with his own account (Figure 6.5(1), the registration interfaces is shown in Figure 6.5(2)). Upon receiving the request, the system automatically offers a few nearby taxis to the passenger together with general evaluations for each of them based on the historical rankings from previous passengers (Figure 6.5(3)). Then he presses the destination area on the mobile phone screen for over three seconds to identify where he wants to go (Figure 6.5(4)), and after he chooses one taxi and issues the delivery request (Figure 6.5(5)). The server directs the request to the taxi client, who can choose to accept or decline the request (Figure 6.5(6)). If rejects, the taxi driver may explain the reasons with interface Figure 6.5(7). Otherwise, both the passenger and driver clients will enter the waiting state for the driver to come to pick the passenger up. During this process, the passenger can see the location of the taxi to know where it is in real time (Figure 6.5(8)) to alleviate the anxiety of waiting. After picking up the passenger, the system reveals the normal routes (Figure 6.5(9)) to both sides and in case the driver follows an anomalous routes, the system can provide real time alert to the passenger if he wants. As an illustrative and simple design, the delivery route is marked blue when it's normal and marked red when anomaly occurs (Figure 6.5(10)). When the delivery is finished, a summary of the trip is revealed on the screen of both clients (Figure 6.5(11)).

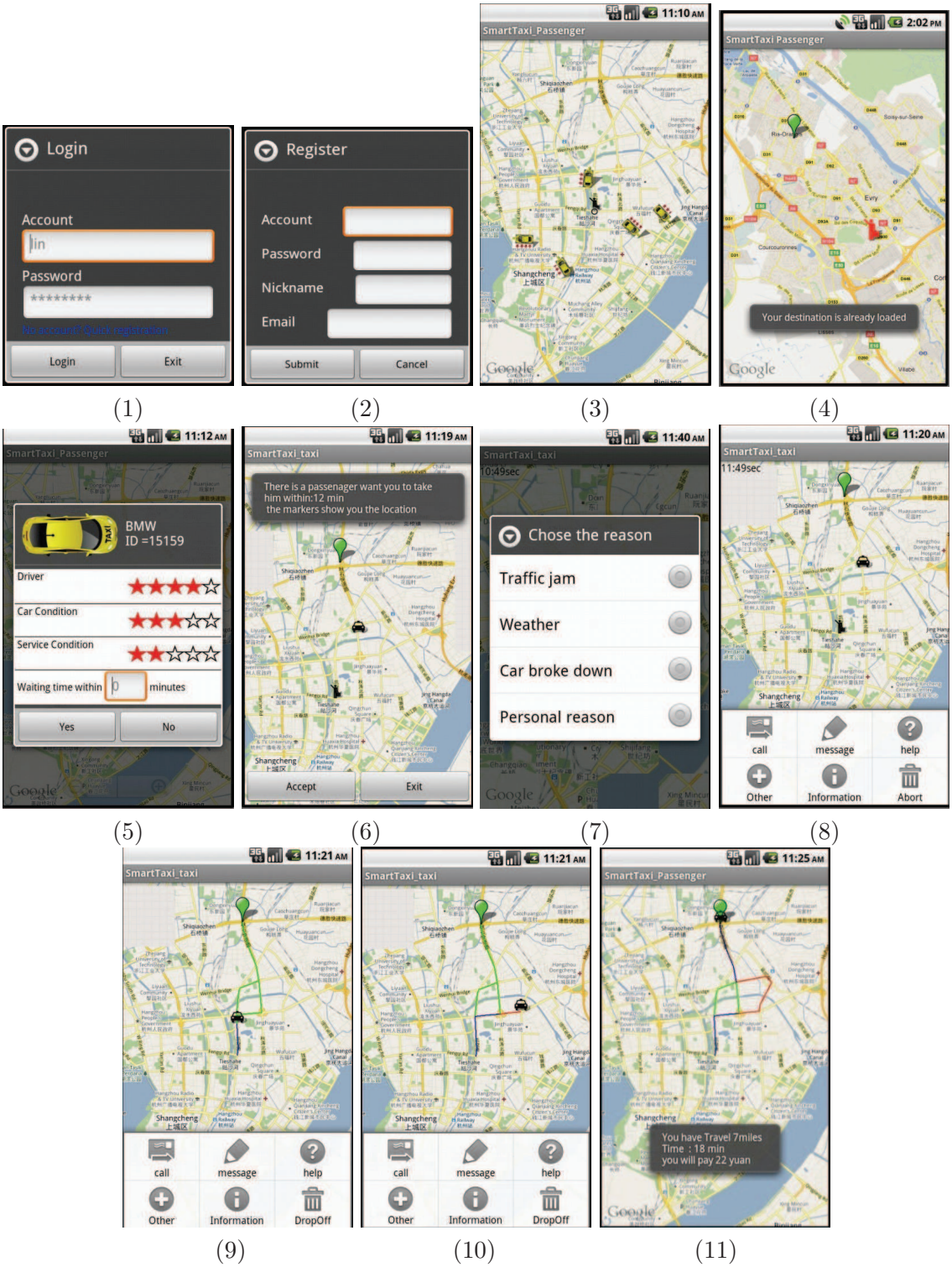


Figure 6.5: Mobile phone user interfaces.

6.2.4.2 Anomalous Passenger Delivery Monitor

Anomalous Passenger Delivery Monitor presents various aspects about the anomalous passenger delivery behaviors in a city in a real time manner to the taxi companies. Currently it mainly includes three interfaces, *i.e.*, real-time monitoring the anomalous passenger delivery behaviors in a city, ranking the drivers according to their tendency to detours (based on criterion such as detour frequency, detour times or detour distance), and searching historical detours. Figure 6.6 shows the three snapshots for the functions listed above respectively. In Figure 6.6(a) the red taxis are those who are conducting anomalous behaviors at the moment. In Figure 6.6(b) the user can select time period and ranking criteria, and then the system will provide a ranking of taxis. The user can choose a taxi and visualize the detour trajectories. In Figure 6.6(c), we provide a searching interface, which given a taxi ID and time period, provides the anomalous trajectories satisfy the criterion.

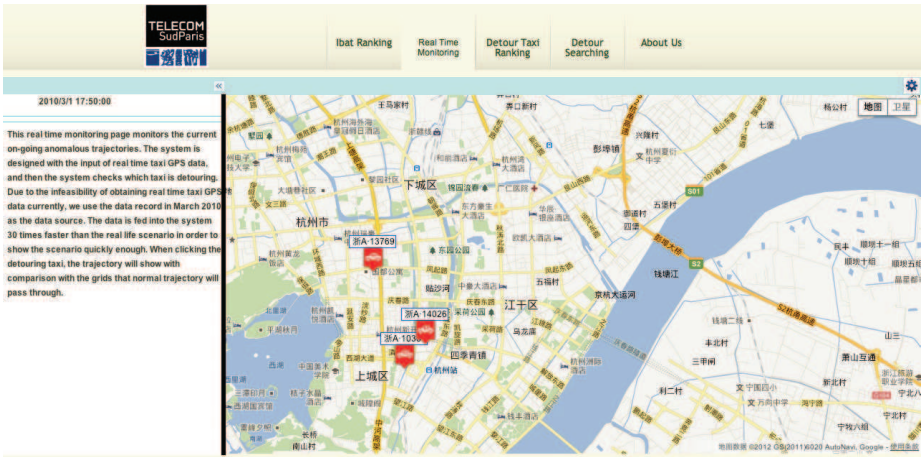
6.2.4.3 Estimating Traveling Time

The traveling time from a source to a destination is highly influenced by the chosen route and its real time traffic. In practice, we can simply estimate it by averaging the historical traveling time of similar trips between these two areas in the same contexts (time of day, day of week and weather) [5]. It triggers the process of extracting a trajectory dataset with same source and destination areas, which is the same process as the anomalous passenger delivery detection work (both *iBAT* and *iBOAT*). We are going to evaluate this aspect in real life cases.

6.2.4.4 Road Network Change Detection

The road network change here includes the blocked or newly-built road segment, which if we look from the vehicle trace perspective, causes the vanishing or emerging of travel patterns in the corresponding places respectively. Cao et al. [13] proposed a method to build a routable road map from the collected vehicle GPS traces. Here we present a easy complementary method to isolate the road network change areas, whereafter the road map construction task focuses only on the network changes.

We first use grid decomposition to partition the city into a grid cell matrix (the dashed grids in Figure 6.7). A grid cell has 8 orientations as shown in the grid decomposition in Figure 4.5. For each grid cell, we count the visiting frequency, *i.e.*, the number of trips that traverse through it, in each orientation. For a block road segment, as illustrated in Figure 6.7(a), the visiting frequencies of the grids along with the road direction, which cover the changed road segment will change from a big number to zero. On the contrary, they change from zero to a big number for a new road segment (shown in Figure 6.7(b)). So



(a)

(b)

(c)

Figure 6.6: Smart Taxi Monitor - User interface for administrators in web browser.

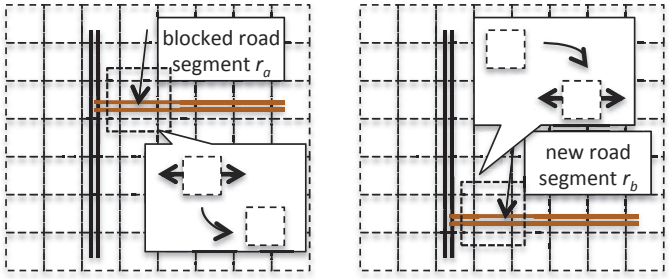


Figure 6.7: Examples of isolating road network change. (a) blocked road segment; (b) new road segment.

by checking the visiting frequency change, we can easily isolate the road network change.

6.3 System Evaluation

We are building a prototype of the smart taxi system. In this section, we evaluate our system with the collected taxis' GPS dataset. As the anomalous passenger delivery behavior monitoring requires real time responses, we evaluate the system response time to show it's effectiveness in dealing with large number of taxis. Meanwhile, as both the detection method and traveling time estimation method require a dataset with sufficient number of trajectories under the same $\langle S, D \rangle$ pair, we also evaluate this assumption in real life scenarios. We define the system coverage as how much percentage of trajectories in real life that fulfill such a requirement and evaluate it in different situations.

Currently there are about 8,400 taxis operating in Hangzhou, China, out of which around 7,600 have been deployed with a GPS sensing device. In March 2010 we collected about 441 million packets from these taxis. In this section we intend to verify whether our proposed system can handle anomaly detection for such a large fleet of taxis. For the time settings such as different time periods of day and different day of week, we choose the same one as in Section 5.1.

We calculate the average daily delivery trips for each taxi and show the distribution over these averages in Figure 6.8. It shows that most of the taxis have about $30 \sim 50$ trips per day, but there are a few excellent taxis that make over 50 trips per day. We display the average number of daily trips in each time slot in Table 6.1. Although there are more trips in the day time of working days, there are more trips in night time of non-working days. This behaviour is in accordance with what is expected from our experience, as in working days, people go out more often at day time, while in non-working days, people go out more

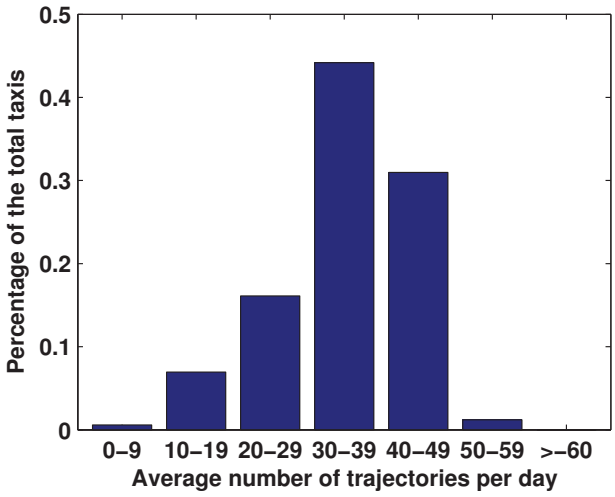


Figure 6.8: Distribution over average number of daily trips.

often at night time for entertainment and so on.

Table 6.1: Average number of daily trips

Day type	Night 0∼6:59	Morning 7:00∼11:59	Afternoon 12:00∼16:59	Evening 17:00∼23:59
Working	27k	58k	63k	87k
Non-Working	34k	54k	61k	95k

6.3.1 System Response Time

We would like to verify that our system is able to handle data coming from a large fleet of taxis and monitors the anomalousness of on-going trajectories in a timely manner. Specifically, since the sampling rate of GPS devices is approximately once per minute, we would like to ensure that our system can respond to all taxis within one minute.

We construct our database using data from March 2010, and use the data from April 2010 to simulate a deployed system. Since each GPS packet includes a timestamp, we can "replay" each of these packets in the correct order and at the correct time, effectively replicating the reception of these packets in the real life scenario. We ran the test on a PC with Intel Xeon CPU (2.8GHz) and 12G memory.

For the dataset of 7,600 taxis, our system was able to respond within 2 seconds, which is well below one minute. We envision this type of system can be deployed in larger cities, such as Beijing (around 70,000 taxis) and Mexico city (around 250,000 taxis), so it is

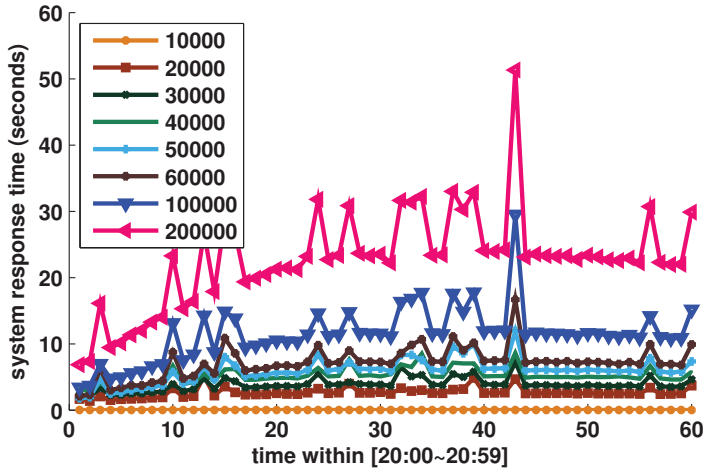


Figure 6.9: Response time versus number of taxis in the busiest hour.

necessary to verify the scalability of the proposed system. We "cloned" the taxi records in our database in order to create a large fleet of taxis. In Figure 6.9 we display the response time of the system during the busiest hour in our dataset (20:00∼21:00). We can see that the response time grows no more than linearly with increasing data sizes, remaining well below a minute even for a fleet of 200,000 taxis. This suggests that even for a city with around 200,000 taxis, the response time should be below one minute by deploying the system in an office PC.

6.3.2 System Coverage with Extracting Similar Trajectories

The detection of anomalous points is based on the number of *support* from historical trajectories. If there are very few trajectories between a particular source and destination area, the method will not be able to check incoming trajectories between said pair of points. Thus, given a fixed dataset of trajectories, the threshold we select for determining the sufficiency of trajectories directly determines how many source-destination pairs can be tested by *iBOAT* and how much percentage of trajectories can be tested by our system (*i.e.*, system coverage) in real life. We choose the grid size $500m \times 500m$, which is the same as in [125]. In Figure 6.10 we plot the percentage of trajectories covered by our system as the threshold varies with all the trajectories in March 2010. We can see it drops when the threshold increases. As there are 23 working days in that month while only 8 non-working days, the coverage in each time slot of day is bigger in working days than in non-working days. As the criterion of judging whether the trajectories are anomalous is decided by whether it is "few" and "different" from the majority, in other words, as long as

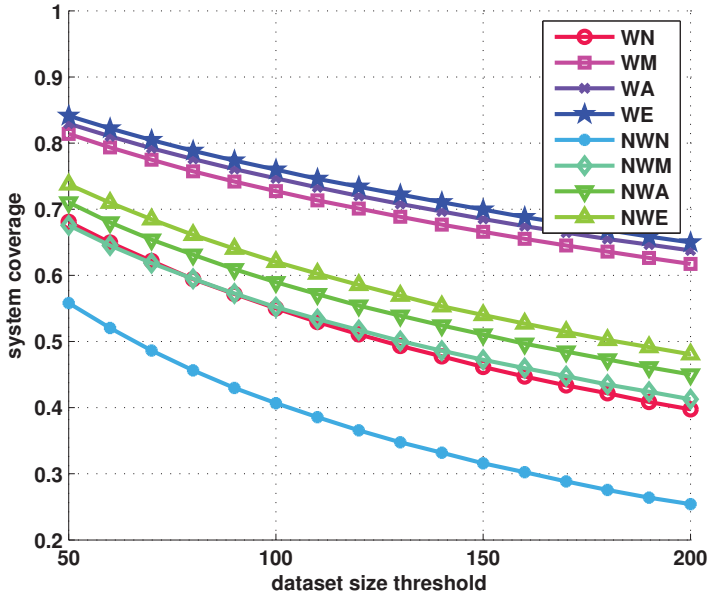


Figure 6.10: System coverage versus sufficient trajectory dataset size thresholds (Using 4 weeks historical data).

the "majority" of trajectories are normal routes, the anomaly detection method will work. We visualize many source and destination paris, and find that 50 is a quite safe threshold. It indicates that at least 55% of the trajectories will be covered during the night time of non-working day, and over 80% during the morning, afternoon and evening time of working days.

One way to increase the system coverage is to increase the time period of historical dataset, which means that we consider more amount of historical information. In Figure 6.11 we vary the number of weeks used to build the historical database and display the percentage of trajectories covered when the threshold is fixed at 50. We can see that with more weeks the system coverage increases but the increment per week is decreasing. It's because, with longer historical data, the uncovered pairs are those with fewer trajectories in one week and thus the increment is slow.

Another way is to increase the size of source and destination area, which will include more trajectories eventually. The detected anomalies are the delivery routes that are not complying with the normal routes between the $\langle S, D \rangle$ areas. With bigger areas, the system gathers more trajectories between two areas and then some anomalies may become normal and can't be detected in our system. But still, the detected anomalies are correct as they are "few" and "different" from the normal routes. We evaluate the system coverage with

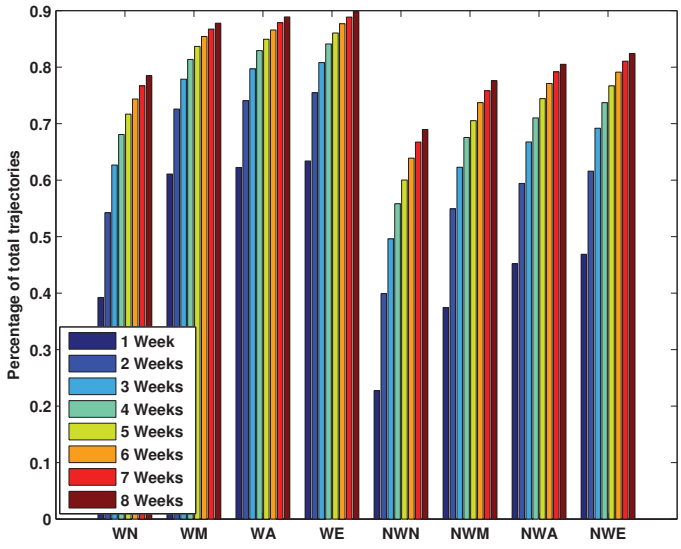


Figure 6.11: System coverage versus number of weeks of historical data (Threshold=50).

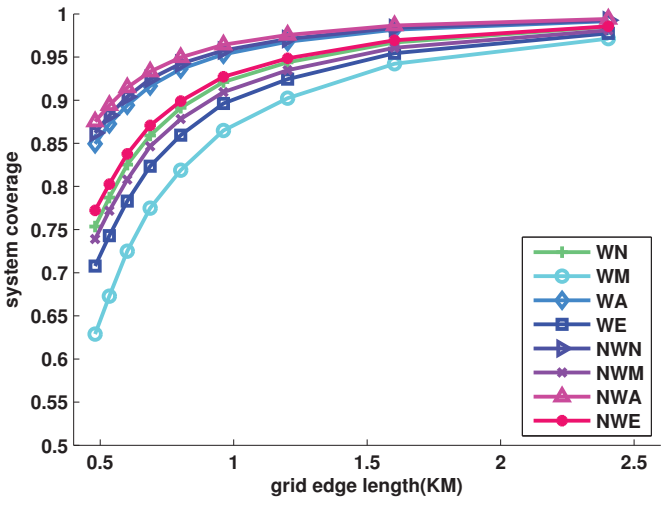


Figure 6.12: System coverage versus source/destination grid size (Threshold=50, one month data)

respect to different area size (grid edge length) and show the result in Figure 6.12. We can see with one month taxi data and threshold 50, in working days, the coverage increases to over 95% when we choose $1km \times 1km$ grids, while in non-working days, it achieves 90% with $1.5km \times 1.5km$ grids.

6.4 Conclusion

Enriched by the perception and understanding of individual and community behaviors from the digital footprints, the service scope and functionalities of the smart sensing systems are greatly expanded. We present an exemplar study with a smart taxi system in this chapter, which doesn't merely provide direct dispatch and navigation services for drivers and passengers, but a complete and considerate taxi service system that allows passenger to easily and comfortably enjoy taxi services and assist taxi drivers to efficiently serve passengers. Besides, it also adopts the fruits born from the exploration of the digital footprints, especially the monitoring and understanding of human behaviors, and supports various types of value-added services to possible clients other than taxi drivers and passengers.

We identify three types of clients who potentially benefit from the services provided by this smart taxi system. They are the smart taxis, smart passengers and smart monitors. Specifically, a central server is deployed in the middle of the system, receiving the GPS reports from taxis and providing services to all the three clients. It provides passenger finding guidance, taxi dispatch, navigation and smart payment services to smart taxis and easy-to-use taxi order services to smart passengers. Besides, it provides various services to other clients like the taxi company, city traffic bureau and even the public, such as monitoring real time anomalous behaviors and traffic, revealing the hotspots and human mobility patterns and so on. We present a system architecture of the central server which consists of four modules an elaborate the design of them separately.

We evaluate the prototype system by "replaying" the collected taxi GPS data with the right time order. Based on our proposed trajectory anomaly detection method, the system response time is within 1 minute with even 200,000 taxis, which is less than the sampling time of a taxi and good enough for real time applications in big cities like Beijing and Shanghai. Besides, we evaluate the system coverage of the methods that requires a taxi trajectory dataset with similar source and destination areas to introduce their usage in real time usage and propose two ways to increase the system coverage.

Chapter 7

Conclusion and Future Work

Contents

7.1 Conclusion

With the prevailing of smart sensing systems in our daily lives, such as the mobile phones and GPS navigators, we are able to obtain people's digital footprints left during their usage of such devices. The large collections of such digital traces provide unprecedented opportunities for us to study people's individual behaviors and community behaviors. In this thesis, we conduct research to study one's behaviors based on his individual digital footprints, and to understand human behaviors from a community perspective, investigating their characteristic and uncovering the hidden human intelligence. The studies are based on the accelerometer digital footprints of mobile phones when they are placed in people's daily costume pockets and the GPS traces of a large taxi fleet. The challenges are mainly from three aspects. Firstly, the coarse quality of the digital footprints normally need data refinement to satisfy the requirement of the target research. Take the GPS traces of taxis as an example, there are entries missing or erroneous due to device failure, network failure or GPS blind spots. Besides, there are no clear indications of the drivers and the taxis may be in sleep mode as the drivers need to park and sleep. Secondly, there are big gaps to be bridged between the low level sensing records and the high level representation of human

behaviors. It's always difficult to directly tell the behaviors from the sensing signals. For example, we have to extract features from the accelerometer sensing data to present the characteristics that potentially can differentiate different physical activities and then use data mining methods (such as SVM, decision tree and so on) to train a model and assign labels to new samples with the model. Thirdly, with the models of the individual behaviors, how to learn the hidden human intelligences, such as evaluating the good and inappropriate behaviors, is a big challenge.

This thesis studies the individual and community behaviors hidden in the digital footprints of people's mobile phone and taxi GPS devices. Specially, we seek the opportunities when people put their mobile phones in the pockets around the pelvic region to recognize their physical activities. We use the accelerometer sensing of mobile phones in different pocket locations with natural orientations to collect the digital footprints. We segment the accelerometer sensing data into half overlapping windows and extract the features including the mean, variance, correlation, DFT energy and entropy. Then we build a the feature matrix and classify the activities with different data mining methods. Noticing that the acceleration magnitude is independent of the mobile phone gestures, we add it to the sensor readings and successfully increase the recognition accuracy. Also we propose a simple method to get rid of the sensing attributes with little contribution and obtain a compact model with little loss of the recognition accuracy. With the digital traces of taxis, we study the anomalous passenger delivery behaviors, summarizing them into three categories with their unique characteristics that can be used for the anomaly detection. We introduce the anomaly detection method *iBOAT* and present our effort in improving its efficiency. By adopting an inverted index mechanism, we successfully improve the efficiency of *iBOAT* at least 5 times faster. We also detect the work shifting events of taxis and separate the digital traces for each individual taxi driver. Based on the observation that work shiftings of a taxi normally happens in a fixed location within a fixed time period of day, and takes some time to handle over the vehicle, we detect the work shifting events in the following two steps. Firstly we find the work shifting location candidates by checking which area the taxi goes to and stays routinely each day. Then we filter out the false candidates by rules obtained in the observation. With the work shifting events, we successfully obtain the digital traces for each taxi driver.

With a large collection of digital traces from thousands of taxis, we obtain a huge amount of anomalous passenger delivery behaviors and conduct thorough analysis to understand them, extract their common characteristics, uncover their motivations and investigate their influences. We find that: 1) over 60% of the anomalous trajectories are "detours" that travel longer distances and time than normal trajectories; 2) the average trip length of drivers with high-detour tendency is 20% longer than that of normal drivers;

3) the length of anomalous sub-trajectories is usually less than a third of the entire trip, and they tend to begin in the first two thirds of the journey; 4) although longer distance results in a greater taxi fare, a higher tendency to take anomalous detours does not result in higher monthly revenue; 5) taxis with a higher income usually spend less time finding new passengers and deliver them in faster speed. We also analyze the work shifting events from a community perspective and validate that, work shifting locations are normally in non-hot areas and the afternoon work shiftings normally happen within 16:40∼17:20, which is why people feel hard to get taxis around that time period.

By modeling and mining the behaviors of different taxi drivers, we intend to understand the taxi serving strategies and uncover those good and bad ones. While modeling the passenger finding behaviors of a driver, we have to extract the initial decisions they make right after the drop-off events. This thesis proposes a novel method capturing that inefficient passenger finding routes normally contain decision making processes in the middle. So we propose a novel method to extract the longest normal sub-trajectory in a passenger finding trajectory and treat it as the initial decision process. After, we model the passenger finding and delivery strategies as well as the passenger serving areas of all drivers and form a matrix which describes their preferences over each of the strategy. We propose methods to uncover the good and bad strategies by learning from the good drivers, measuring their correlations with the taxi performance and finding the strategies that differentiates good and bad taxi drivers.

The individual and community behaviors extracted from the digital footprints enables novel applications of smart sensing systems. This thesis demonstrates it with a smart taxi system, which serves as smart agents for both passengers and taxi drivers, and provides extra services to the public, such as monitoring the traffic and human mobilities, estimating traveling time and providing anomalous alerts. We present the roles of three system clients, including the smart taxi, smart passenger and smart monitor and the system architecture design. The evaluation of the prototype system proves that it supports real time responses. And also we reveal the system applicability of the anomalous passenger delivery detection method.

In summary, the contributions of this thesis are:

1. By mining the accelerometer footprints collected when people place their mobile phone inside the pockets of daily costumes, we successfully recognize seven daily physical activities. We conduct real life experiments and prove that, we do able to detect people's physical activities and by introducing the acceleration magnitude into the sensor reading, the recognition accuracy could be improved about 8%. The cross validation results of several data mining methods show that, SVM achieves the best accuracy. With a simple feature reduction mechanism, we successfully reduce the

computing cost.

2. We summarize the anomalous passenger delivery behaviors, pinpointing their unique characters and the suitable methods to detect them. And then we introduce an inverted index mechanism to replace the trajectory searching in *iBOAT* with index comparisons. The results show that, it improves the computing efficiency at least 5 times.

3. We detect the work shifting events based on the digital footprints of taxis for the first time. By interviewing with several taxi drivers, we find that the two drivers serving one taxi normally have an agreed work shifting location and time period and spend some time for the handover of vehicles. So we extract the waiting locations in the vacant trajectories and map them into grid decomposition. The grids that the taxi stays nearly everyday in a fixed time slot are counted as the possible candidates. After we propose rules to filter out the false candidates and obtain the real work shifting location. And then we detect the work-shifting events and obtain the GPS traces for individual drivers.

4. We detect huge number of anomalous passenger delivery behaviors from the digital traces of a taxi fleet and provide thorough analysis of their characteristics and influences on taxi revenues. We reveal that, these anomalous behaviors normally cost longer traveling distance and more traveling time, which means that most of them are detour events. Even though high detour tendency taxis earn more averagely in single trips, they are not the top revenue taxis. Top revenue taxis are good at efficient passenger finding and passenger delivering, and don't rely on the fraud behaviors.

5. We uncover the good and bad taxi serving techniques from the digital footprints of a community of drivers. We first model the taxi serving behaviors of each driver based on his digital traces from the perspectives of passenger finding, passenger delivering and serving areas. Particularly, we study the passenger finding intentions right after dropping off passenger for the first time and propose a novel method to extract the intentions from the passenger finding trajectories. The taxi serving behaviors of a driver are described as a feature vector, which reveal one's preferences over different strategies. Then we study the good and bad strategies from the perspectives of learning from top driver, measuring by the correlation between each strategy preference and the performance and finding the strategies that differentiate the drivers. We present the methods and results accordingly.

6. Furthermore, we present a smart taxi system, which explores the individual and community behaviors mined from the digital traces to support various novel applications. We present the possible users, system architecture and evaluations in real life

scenarios.

7.2 Discussions and Future Work

7.2.1 Understanding Individual Behaviors

Recognizing and understanding individual behaviors from the low level sensing data will greatly increase the intelligence level of our computing systems in the future. With the understanding of an individual's behaviors, the system can adapt its own behaviors to cater for the needs of the user, provide assistance, and prevent wrong behaviors that may harm people. Such motives already inspired many researchers to devise various ways to recognize human behaviors with sensors. However, there are still a long way to go due to the following two challenges.

One is the high cost of obtaining labeling data. Traditional supervised learning methods based activity recognition require labeled data so that it can learn models to differentiate the classes. However, practically, obtaining the labeled data is time and effort costly. It normally relies on manually label work from either the testers or observers to tag the data (in our case, we design a touch screen interface to help the volunteers to tag the data. But it still cost them lots of effort.) As pointed out in [99], these methods have significant deficiencies in cost, accuracy, scope, coverage and obtrusiveness. Extensive observation causes fatigue in observers and resentment in those being observed. In addition, the constant human involvement makes the process costly. Self-reporting is often inaccurate and of limited usefulness due to exhaustive patience and intentional and unintentional misreporting.

The solutions to this problem may from the following aspects. The first one is to design user-friendly interfaces to ease the burden of labeling data. For example, to label people's physical activities, we designed a touch-screen interface that is easy to use. Imagining that by adopting voice recognition or some other technologies that are mature enough, we are expecting new ways to label the data. The second one is to make use of the unlabeled data to recognize the activities, such as adopting semi-supervised learning methods or even unsupervised learning methods. Semi-supervised learning is a type of machine learning techniques that make use of both labeled and unlabeled data for training typically a small amount of labeled dat with a large mount of unlabeled data [16]. For example, Longstaff et al. [69] evaluate ways to augment the activity recognition accuracy with unlabeled data and prove that, when the initial classifier's accuracy is low, active learning [46], En-Co-Training [33] and democratic co-learning [131] can improve the recognition performance greatly. But when the initial accuracy is high, they are useless but also harmless. Wyatt et al. [111] treat activity data as a stream of natural language terms and build models to

map these terms to activity names. They prove that, in such a way they can achieve an accuracy of 42% over 26 activities with unlabeled RFID sensing data.

The other challenge is the extreme diversity of both the contextual conditions and user characteristics encountered in the real-world. Take the recognition of physical activities with mobile phones for example. People carry the mobile phones in many different ways, such as holding in their hands, placing in their pockets and bags, or even in the car or on a table. This thesis proved that we can recognize people's physical activities when their mobile phones are placed in daily costume pockets. But such a model will certainly fail if testing with mobile phones in their hands, or on a table. Because the movement of the hands are different with the pelvic regions, not to mention the scenario on a table. This problem also happens in the taxi GPS traces. For example, the anomalous behaviors may be caused by intentional detours, or by blocked routes due to reasons like traffic accidents, or by some other subjective reasons. So it's hard to exactly tell whether they are conducting fraudulent behaviors merely based on the GPS traces. The diversity of user characteristics are also a big problem. Even we fix the location of the mobile phone, the gait of different people may be quite different. So the model obtained from the labeled data of a few individuals will probably fail if applied to the others. Meanwhile, the acquisition of large scale training data for each user is hardly practical in real life. In his Ph.D thesis recently defensed, Nicholas D. Lane referred to it as the *population diversity problem*. People can have different cultures, live in different places, be different in height, weight, sex and which extent they are physically active. So they may do the same activity in different ways, which cause great diversity in the modeling of the behaviors. Lane et al. propose to solve these problems by looking at people from the community perspective instead of individually. They propose a framework called *CoCo* which leverages different types of everyday social connections between people to build personalized classification models [55]. It exploits social networks to selectively combine small contribution of labeled data from people with shared context or user characteristics to relieve the population diversity problem.

7.2.2 Mining Community Digital Traces

The large collections of digital footprints are just emerging in our live. We believe that, in the near future, the large aggregation of digital footprints in people's mobile phones (sensing data + people's interactions with the mobile phones), the radio signals received in the cell tower, the GPS traces of vehicles and people, the ticketing data in public transportation, the user generated contents (tweets, check in, photos), transportation sensor network (cameras and loop sensors, parking lots), environmental sensor network (air quality, temperature, radiation), transaction records of credit cards, shopping records and so on, will bring increasingly deep understandings of human beings and the society and provide

smart and considerate services to people, local communities and even the whole society in various aspects ranging from personal living and working to community management, criminal investigation, city dynamics and etc. Zhang et al. [124] propose *Social and Community Intelligence* which goes beyond a signal smart space to the community level and identify three main data sources which are multimodal and heterogeneous, including

1. Social network and Internet interaction services, which provide data about individual's preferences and social relationships;
2. infrastructure-bound sensor data about environment;
3. mobile and wearable sensor data about the individual and moving objects.

The mining of these sources together supports services from social networks, urban sensing, environment monitoring, public safety, amongst others.

A recent event in China implies the great potentials of investigating the community digital footprints in social management and criminal detection. In the strikes recently happened all over China against Japan's illegal "purchase" of the *Diaoyu Island*, which is part of China's territory, a few criminals hidden in the strike destroyed many Japanese-brand vehicles (both public and personal belongings) without fear of law because it's hard to identify them from the crowded group and "laws are not laid down to punish the majorities". An extreme case was happened in Xi'an, the capital city of Shaanxi province, Jianli Li, owner of a "Toyota" private car, was severely attacked with his skull broken by one criminal while protecting his vehicle. Traditionally this case is hard to solve as the criminals are hard to be isolated from the whole strike group with clear evidence. However, with the video surveillance recordings on the street, the police obtained a picture of the criminal when he was destroying a vehicle and knew that he wore a T-shirt with a character "D" (shown in Figure 7.1(a)). Afterwards published on the news, many citizens reported the photos of the crime scene they took with their mobile phones, and the police obtained a number of clear pictures of the criminal (such as the one shown in Figure 7.1(b)), which provide evidences and greatly help to catch him.

This case shows the great power with the primitive usage of the digital footprints (the video/camera records in this case) in our daily lives. Even though this case relies much on people's participation, we can design sophisticate process to automate and promote the process with people's digital footprints. For example, as people normally carry their mobile phones, with the radio signal traces collected in the nearby cell towers, we are able to roughly locate those ones around the crime scene and limit the investigation range. With the bluetooth traces, we are able know who was close to whom during the process and hopefully we can reduce the investigation range again by finding people around the vehicle and identifying those around them. And also with the verification of the videos and photos recorded by the street camera and people's mobile phones, we hope to identify the criminal

(a) (b)

Figure 7.1: Photos of the criminal taken by a street video surveillance system (a) and people's mobile phones (b).

quickly. By mining from the photos and videos, the sensing traces of the criminal's mobile phones, we expect ways to build concrete evidence to prove the crime he committed.

There are several research issues that arise in the above scenarios, which can motive much research in the future. The first one is how to deal with the large-scale heterogeneous data sources. Different types of sensors have different attributes, qualities and capabilities. And they may be produced in different locations by different individuals (cameras in the street and personal mobile phones), which may cause synchronization problem and bring trouble to the learning process. Besides, the system often deals with huge amount of data collected from large number of individuals and suffers the computational difficulties. So much work are needed on sampling optimization, problem decomposition, model optimization within particular problem domains. The second one is the privacy and trust problem, although it's not evolved in this thesis. As we are using the data from many individuals, we have to protect their privacy so that people can trust us and share their digital footprints which potentially expose their privacy. As in our case, the police can act with juridical rights given by the law. However, in many cases, we don't have such rights and have to carefully protect people's privacies and build trust among them for the system. There are two possible solutions currently, data anonymization and user control, which tries to protect the contents and the access to the contents. The third one is the sensing mechanism. Should it be participatory sensing or opportunity sensing, or a tradeoff between them? In participatory sensing, people decide whether they perform the sensing tasks (such as, whether to take photos of the crime scene and share to the police), while in opportunistic sensing, the system automatically decides when to use devices to meet the application's

sensing requests [54].

My future work will follow the way to exploit the digital footprints produced by the sensing devices in our life to devise novel services to individual people, local communities and even our society. The above listed problems are the possible future research topics.

Bibliography

[1] F. Allen, E. Ambikairajah, N. H. Lovell, and B. G. Celler. Classification of a known sequence of motions and postures from accelerometry data using adapted gaussian mixture models. *Physiological Measurement*, 27(10):935, 2006.

[2] S. Amarnag and R. Alvin. Recognizing activities and spatial context using wearable sensors. In *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[3] L. Atallah, B. Lo, R. Ali, R. King, and G.-Z. Yang. Real-time activity classification using ambient and wearable sensors. *Trans. Info. Tech. Biomed.*, 13(6):1031–1039, 2009.

[4] L. Atallah and G.-Z. Yang. Review: The use of pervasive sensing for behaviour profiling - a survey. *Pervasive Mobile Computing*, 5(5):447–464, 2009.

[5] K. Balan Rajesh, X. Nguyen Khoa, and J. Lingxiao. Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, pages 99–112, 2011.

[6] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, pages 1–17, 2004.

[7] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3:1183–1208, 2003.

[8] M. Berchtold, M. Budde, D. Gordon, H. Schmidtke, and M. Beigl. Actiserv: Activity recognition service for mobile phones. In *Wearable Computers (ISWC), 2010 International Symposium on*, pages 1–8, 2010.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[10] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[11] Z. D. Brian, A. L. Maas, A. K. Dey, and J. A. Bagnell. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 322–331, 2008.

[12] H. Cao, N. Mamoulis, and D. Cheung. Mining frequent spatio-temporal sequential patterns. In *Fifth IEEE International Conference on Data Mining*, pages 82–89, 2005.

[13] L. Cao and J. Krumm. From gps traces to a routable road map. In *Proc. GIS 2009*, pages 3–12, 2009.

[14] C. C. Chang and C. J. Lin. *LIBSVM: a Library for Support Vector Machines*, 2001.

[15] H. Chang, Y. Tai, and J. Y. Hsu. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, 5:3–18, 2010.

[16] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2010.

[17] C. Chen, D. Zhang, P. S. Castro, N. Li, L. Sun, and S. Li. Real-time Detection of Anomalous Taxi Trajectories from incoming GPS Traces. In *Proceedings of the 8th International ICST Conference on Mobile and Ubiquitous Systems*, 2011.

[18] G. Chen, X. Jin, and J. Yang. Study on spatial and temporal mobility pattern of urban taxi services. In *2010 International Conference on Intelligent Systems and Knowledge Engineering*, pages 422–425, 2010.

[19] L. Chen, M. Lv, and G. Chen. A system for destination and future route prediction based on trajectory mining. *Pervasive Mobile Computing*, 6(6):657–676, 2010.

[20] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. P. Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, 7(2):32–41, 2008.

[21] J. Christanini and J. Taylor. *An Introduction to Support Vector Machines and other Kernelbased Methods*. Cambridge University Press, 2000.

[22] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the 26th annual SIGCHI conference on Human factors in computing systems*, pages 1797–1806, 2008.

[23] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.

[24] A. R. Doherty, Z. Qiu, C. Foley, H. Lee, C. Gurrin, and A. F. Smeaton. Green multimedia: informing people of their carbon footprint through two simple sensors. In *Proceedings of the international conference on Multimedia*, MM '10, pages 441–450, 2010.

[25] P. Donald J., L. Lin, F. Dieter, and K. Henry A. Inferring high-level behavior from low-level sensors. In *Proceedings of Ubicomp 2003*, pages 73–89, 2003.

[26] S. Edelkamp and S. Schrödl. Route planning and map inference with global positioning traces. In *Computer Science in Perspective*, pages 128–151. Springer-Verlag New York, Inc., 2003.

[27] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):20 –26, 2008.

[28] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou. A taxi driving fraud detection system. In *Proceedings of the IEEE International Conference on Data Mining*, pages 181–190, 2011.

[29] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695–719, 2011.

[30] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag. Adaptive fastest path computation on a road network: a traffic mining approach. In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 794–805, 2007.

[31] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[32] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu. epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition. In *Proceedings of the 2009 IEEE International Conference on Pervasive Computing and Communications*, pages 1–9, 2009.

[33] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee. Activity recognition based on semi-supervised learning. In *Proceedings of the 13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, RTCSA '07, pages 469–475, 2007.

[34] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11:10–18, 2009.

[36] T. Huỳnh, U. Blanke, and B. Schiele. Scalable recognition of daily activities with wearable sensors. In *Location- and Context-Awareness*, pages 50–67. Springer Berlin / Heidelberg, 2007.

[37] Y. Hyoseok, Y. Zheng, X. Xie, and W. Woontack. Smart itinerary recommendation based on user-generated gps trajectories. In *Ubiquitous Intelligence and Computing*, volume 6406, pages 19–34, 2010.

[38] F. Ichikawa. Where is the phone? a study of mobile phone location in public spaces. In *Proceedings of the International Conference on Mobile Technology, Applications, and Systems*, pages 797–804, 2005.

[39] S. S. Intille, K. Larson, E. M. Tapia, J. S. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. Using a live-in laboratory for ubiquitous computing research. In *Proceedings of the 4th international conference on Pervasive Computing*, pages 349–365, 2006.

[40] S. Isaacman, R. Becker, R. Cáeres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, 6696:133–151, 2011.

[41] T. Iso and K. Yamazaki. Gait analyzer based on a cell phone with a single three-axis accelerometer. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 141–144, 2006.

[42] X. Ji and H. Liu. Advances in view-invariant human motion analysis: a review. *Trans. Sys. Man Cyber Part C*, 40(1):13–24, 2010.

[43] B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80:021136, 2009.

[44] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[45] F. Jon and K. John. Route prediction from trip observations. In *Intelligent Vehicle Initiative (IVI) Technology Advanced Controls and Navigation Systems, SAE World Congress & Exhibition*, 2008.

[46] A. Kapoor and E. Horvitz. Experience sampling for building predictive user models: a comparative study. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 657–666, 2008.

[47] N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *Ambient Intelligence*, volume 2875, pages 220–232. Springer Paris, 2003.

[48] C. Kidd, R. Orr, G. Abowd, C. Atkeson, I. Essa, B. MacIntyre, E. Mynatt, T. Starner, and W. Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *Cooperative Buildings. Integrating Information, Organizations, and Architecture*, volume 1670, pages 191–198. Springer Berlin / Heidelberg, 1999.

[49] J. Krumm and E. Horvitz. Predestination: Inferring destinations from partial trajectories. In *Proceedings of UbiComp*, volume 4206, pages 243–260, 2006.

[50] K. Kunze, M. Barry, E. A. Heinz, P. Lukowicz, P. Lukowicz, D. Majoe, and J. Gutknecht. Towards recognizing tai chi - an initial experiment using wearable sensors. *Applied Wearable Computing (IFAWC), 2006 3rd International Forum on*, pages 1 –6, 2006.

[51] J. Kwapisz, G. M. Weiss, and S. Moore. Activity recognition using cell phone accelerometers. In *Proceedings of the 4th International Workshop on Knowledge Discovery from Sensor Data*, pages 10–18, 2010.

[52] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.*, 12(2):74–82, 2011.

[53] K. V. Laerhoven and H.-W. Gellersen. Spine versus porcupine: A study in distributed wearable activity recognition. In *Proceedings of the Eighth International Symposium on Wearable Computers*, pages 142–149, 2004.

[54] N. Lane, S. Eisenman, M. Musolesi, E. Miluzzo, and A. Campbell. Urban sensing systems: opportunistic or participatory? In *Proceedings of the 9th workshop on Mobile computing systems and applications*, pages 11–16. ACM, 2008.

[55] N. Lane, Y. Xu, H. Lu, S. Eisenman, T. Choudhury, and A. Campbell. Cooperative communities (coco): Exploiting social networks for large-scale modeling of human behavior. *Pervasive Computing, IEEE*, PP(99), 2011.

[56] J. Lee, I. Shin, and G.-L. Park. Analysis of the passenger pick-up pattern for taxi location recommendation. In *Proceedings of the 2008 Fourth International Conference on Networked Computing and Advanced Information Management*, pages 199–204, 2008.

[57] J. Lester, T. Choudhury, and G. Borriello. A practical approach to recognizing physical activities. In *Pervasive*, volume 3968, pages 1–16. Springer, 2006.

[58] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 766–772, 2005.

[59] J. Letchner, J. Krumm, and E. Horvitz. Trip router with individualized preferences (trip): incorporating personalization into route planning. In *Proceedings of the 18th*

conference on Innovative applications of artificial intelligence - Volume 2, IAAI'06, pages 1795–1800, 2006.

[60] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang. Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 63–68, 2011.

[61] Q. Li, Z. Zheng, B. Yang, and T. Zhang. Hierarchical route planning based on taxi gps-trajectories. In *Proceedings of the 17th International Conference on Geoinformatics*, pages 1–5, 2009.

[62] Q. Li, Z. Zheng, T. Zhang, J. Li, and Z. Wu. Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data. *International Journal of Applied Earth Observation and Geoinformation*, 13(1):110–119, 2011.

[63] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science in China*, 6:111–121, 2012.

[64] L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Building personal maps from gps data. In *In Proceedings of IJCAI Workshop on Modeling Others from Observation*, 2005.

[65] L. Lin, P. Donald J., F. Dieter, and K. Henry. Learning and inferring transportation routines. *Artificial Intelligence*, 171:311 – 331, 2007.

[66] L. Liu, C. Andris, and C. Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34:541–548, 2010.

[67] S. Liu, Y. Liu, L. M. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining*, pages 919–928, 2010.

[68] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille. A long-term evaluation of sensing modalities for activity recognition. In *Proceedings of the 9th international conference on Ubiquitous computing*, pages 483–500, 2007.

[69] B. Longstaff, S. Reddy, and D. Estrin. Improving activity classification for health applications on mobile devices using active and semi-supervised learning. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on-NO PERMISSIONS*, pages 1 –7, 2010.

[70] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate gps trajectories. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 352–361, 2009.

[71] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 165–178, 2009.

[72] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 71–84, 2010.

[73] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Mathematical Statist. Probability*, pages 281–297, 1967.

[74] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *International Workshop on Wearable and Implantable Body Sensor Networks*, pages 113–116, 2006.

[75] D. McIlwraith, J. Pansiot, S. Thiemjarus, B. Lo, and G. Yang. Probabilistic decision level fusion for real-time correlation of ambient and wearable sensors. In *5th International Summer School and Symposium on Medical Devices and Biosensors*, pages 117 –120, 2008.

[76] T. B. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[77] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646, 2009.

[78] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[79] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 863–868, 2008.

[80] J. J. Pan, S. J. Pan, V. W. Zheng, and Q. Yang. Digital wall: A power-efficient solution for location-based data sharing. In *Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications*, PERCOM '08, pages 645–650, 2008.

[81] J. Pärkkä, M. Ermes, P. Korpipää, J. Mäntyjärvi, J. Peltola, and I. Korhonen. Activity classification using realistic data from wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):119–128, 2006.

[82] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. *2012 16th International Symposium on Wearable Computers*, 0:44–51, 2005.

[83] W. Pentney, A.-M. Popescu, S. Wang, H. Kautz, and M. Philipose. Sensor-based understanding of daily life via large-scale use of common sense. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, pages 906–912, 2006.

[84] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti. Taxi-aware map: Identifying and predicting vacant taxis in the city. In *Ambient Intelligence: First International Joint Conference*, pages 86–95, 2010.

[85] R. Poppe. Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.*, 108(1-2):4–18, 2007.

[86] J. W. Powell, Y. Huang, F. Bastani, and M. Ji. Towards reducing taxicab cruising time using spatio-temporal profitability maps. In *12th International Symposium on Spatial and Temporal Databases*, 2011.

[87] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang. Measuring social functions of city regions from large-scale taxi behaviors. In *The 9th IEEE International Conference on Pervasive Computing and Communications, WIP*, pages 384–388, 2011.

[88] Z. Qiu, C. Gurrin, A. Doherty, and A. Smeaton. Term weighting approaches for mining significant locations from personal location logs. In *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, pages 20 –25, 2010.

[89] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*, pages 1541–1546, 2005.

[90] S. Reddy, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Determining transportation mode on mobile phones. In *Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers*, pages 25–28, 2008.

[91] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transaction on Sensor Network*, 6(2):13:1–13:27, 2010.

[92] H. Ryan, H. Aude, A. Pieter, and B. Alexandre. Estimating arterial traffic conditions using sparse probe data. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, pages 929 –936, 2010.

[93] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining Knowledge Discovery*, 2(2):169–194, 1998.

[94] R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner. A traffic information system by means of real-time floating-car data. In *9th World Congress on Intelligent Transport Systems*, 2002.

[95] D. Stefanov, Z. Bien, and W.-C. Bang. The smart house for older persons and persons with physical disabilities: structure, technology arrangements, and perspectives. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 12(2):228 –250, 2004.

[96] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele. Adl recognition based on the combination of rfid and accelerometer sensing. In *Proceedings of Second International Conference on Pervasive Computing Technologies for Healthcare*, pages 258 –263, 2008.

[97] L. Sun, D. Zhang, B. Li, B. Li, and S. Li. Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In *Ubiquitous Intelligence and Computing*, volume 6406, pages 548–562. Springer, 2010.

[98] T. Takayama, K. Matsumoto, A.Kumagai, N. Sato, and Y. Murata. Waiting/cruising location recommendation for efficient taxi business. *International journal of System Applications, Engineering & Development*, 5:224–236, 2011.

[99] C. Tanzeem, P. Matthai, W. Danny, and J. Lester. Towards activity databases: Using sensors and statistical models to summarize people's lives. *IEEE Data Eng. Bull*, 29:2006, 2006.

[100] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers*, pages 1–4, 2007.

[101] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9, 2008.

[102] M. Veloso, S. Phithakkitnukoon, and C. Bento. Urban mobility study using taxi traces. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, pages 23–30, 2011.

[103] P. Viola and M. Jones. Fast and robust classification using asymmetric adaboost and a detector cascade. *Advances in Neural Information Processing Systems*, 2(3):1311–1318, 2002.

[104] J. Ward, P. Lukowicz, G. Troster, and T. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1553 –1567, 2006.

[105] G. M. Weiss and J. W. Lockhart. Identifying user traits by mining smart phone accelerometer data. In *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, pages 61–69, 2011.

[106] H. Wen, Z. Hu, J. Guo, L. Zhu, and J. Sun. Operational analysis on beijing road network during the olympic games. *Journal of Transportation Systems Engineering and Information Technology*, 8(6):32–37, 2008.

[107] C. E. White, D. Bernstein, and A. L. Kornhauser. Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1-6):91 – 108, 2000.

[108] D. Wilson and C. Atkeson. Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors. In *Pervasive Computing*, pages 329–334. Springer Berlin / Heidelberg, 2005.

[109] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. *IEEE Int. Conf. on Computer Vision (ICCV 2007)*, pages 1–8, 2007.

[110] W. H. Wu, A. A. T. Bui, M. A. Batalin, L. K. Au, J. D. Binney, and W. J. Kaiser. Medic: Medical embedded device for individualized care. *Artifical Intelligence in Medicine*, 42(2):137–152, 2008.

[111] D. Wyatt, M. Philipose, and T. Choudhury. Unsupervised activity recognition using automatically mined common sense. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 1*, AAAI'05, pages 21–27, 2005.

[112] X. Xiao, Y. Zheng, Q. Luo, and X. Xie. Finding similar users using category-based location history. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 442–445, 2010.

[113] K. Yamamoto, K. Uesugi, and T. Watanabe. Adaptive routing of cruising taxis by mutual exchange of pathways. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5178, pages 559–566, 2008.

[114] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy. Distributed recognition of human actions using wearable motion sensor networks. *J. Ambient Intell. Smart Environ.*, 1(2):103–115, 2009.

[115] G.-Z. Yang. *Body Sensor Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[116] J. Yang. Toward physical activity diary: motion recognition using simple acceleration features with mobile phones. In *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, pages 1–10, 2009.

[117] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recogn. Lett.*, 29(16):2213–2220, 2008.

[118] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie. Mining individual life pattern based on location history. In *Proceddings of 10th International Conference onMobile Data Management: Systems, Services and Middleware*, pages 1 –10, 2009.

[119] J. Yuan, J. Luo, H. Kautz, and Y. Wu. Mining gps traces and visual words for event classification. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 2–9, 2008.

[120] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *18th SIGKDD conference on Knowledge Discovery and Data Mining*, KDD 2012, 2012.

[121] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 316–324, 2011.

[122] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, and Y. Huang. T-drive: Driving directions based on taxi trajectories. In *Proceedings of the 18th ACM International Conference on Advances in Geographic Information Systems*, pages 99–108, 2010.

[123] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun. Where to find my next passenger? In *Proceedings of the 13th ACM International Conference on Ubiquitous Computing*, pages 109–118, 2011.

[124] D. Zhang, B. Guo, and Z. Yu. The emergence of social and community intelligence. *IEEE Computer*, 44:21–28, July 2011.

[125] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces. In *Proceedings of the 13th ACM International Conference on Ubiquitous Computing*, pages 99–108, 2011.

[126] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Towards mobile intelligence: Learning from gps history data for collaborative recommendation. *Artificial Intelligence*, 184–185(0):17 – 37, 2012.

[127] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma. Geolife2.0: A location-based social networking service. In *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, MDM '09, pages 357–358, 2009.

[128] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 247–256, 2008.

[129] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Proceedings of the 13th ACM International Conference on Ubiquitous Computing*, pages 89–98, 2011.

[130] Y. Zheng and X. Xie. Learning location correlation from gps trajectories. In *Proceedings of the 2010 Eleventh International Conference on Mobile Data Management*, MDM '10, pages 27–32, 2010.

[131] Y. Zhou and S. Goldman. Democratic co-learning. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pages 594–202, 2004.

[132] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Survey*, 38, 2006.

Appendix A

Thesis Publications

Journal

- **Lin Sun**, Daqing Zhang, Chao Chen, Pablo Samuel Castro, Shijian Li, and Zonghui Wang *Real Time Anomalous Trajectory Detection and Analysis*, Mobile Networks and Applications (MONET), accepted, 2012.
- Bin Guo, Daqing Zhang, **Lin Sun**, Zhiwen Yu, Xingshe Zhou. "iCROSS: Towards a Scalable Infrastructure for Cross-Domain Context Management", ACM/Springer Journal of Personal and Ubiquitous Computing (PUC), Springer, 2012 (to appear).

Book chapters

- **Lin SUN**, Bin Guo, and Daqing Zhang. *Tools for Handling Context Semantics*. Digital Home Networking. Wiley, 2011.
- **Lin SUN**, Daqing Zhang, and Chao Chen. *Understanding City Dynamics from Taxi GPS Traces*. Creating Personal, Social and Urban Awareness through Pervasive Computing, IGI Global, 2013.

International Conferences

- **Lin SUN**, Daqing Zhang, and Nan Li *Physical Activity Monitoring with Mobile Phones*, In Proceedings of the 9th international conference on Smart Homes and Health Telematics (ICOST) 2011, pp: 104∼111.
- **Lin SUN**, Daqing Zhang, Bin Li, Bin Guo, and Shijian Li. *Activity Recognition on an Accelerometer Embedded Mobile Phones with Varying Positions and Orientations*. In Proceedings of the 7th international conference on Ubiquitous Intelligence and Computing (UIC'10), 2010, pp: 548∼562.
- Bin Guo, **Lin SUN**, and Daqing Zhang. *The Architecture Design of a Cross-Domain Context Management System*. IEEE PerCom Workshop on Middleware Support for Pervasive Computing (PerWare'10), Germany, Apr. 2, 2010.
- Chao CHEN, Daqing ZHANG, Pablo Samuel Castro, Nan LI, **Lin SUN** and Shijian Li. *Real-time Detection of Anomalous Taxi Trajectories from GPS Traces*, In Proceedings of 8th International ICST Conference on Mobile and Ubiquitous Systems (MobiQuitous'11), Copenhagen, Denmark, 2011.
- Daqing Zhang, Nan Li, Zhi-hua Zhou, Chao Chen, **Lin SUN** and Shijian Li. *iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces*, In Proceeding of the 13th ACM International Conference on Ubiquitous Computing (UbiComp'11), Beijing, China, 2011.
- Bin Li, Daqing Zhang, **Lin SUN**, Chao Chen, Shijian Li, Guande Qi, and Qiang Yang. *Hunting or Waiting? Discovering Passenger-Finding Strategies from a Large-scale Real-world Taxi Dataset*, In Proceedings of IEEE PerCom Workshops 2011, pp:63-68.

List of figures

List of tables