MODELLING INTRA-URBAN HUMAN ACTIVITY PATTERNS USING CROWDSOURCING GPS AND GEOSOCIAL MEDIA DATA

by

Wei Huang

B.Sc., Beijing University of Chemical Technology, Beijing, China, 2008

M.Sc., Central South University, Changsha, China, 2011

A dissertation

presented to Ryerson University

in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in the Program of

Civil Engineering

Toronto, Ontario, Canada, 2016

© Wei Huang 2016

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Modelling Intra-Urban Human Activity Patterns using Crowdsourcing GPS and Geosocial Media Data

Doctor of Philosophy 2016

Wei Huang

Civil Engineering

Ryerson University

Abstract

The way people live in cities forms human activity patterns, which affects how urban systems work. Therefore, it is essential to understand human activity patterns, where precise prediction of human movements and mechanistic modelling of human activity patterns are the two keys. Most of existing work on prediction of human movements cannot deal with activity changes, leading to a negative impact on the predictive accuracy. Furthermore, the majority of current work on modelling human activity patterns are mainly researched from spatiotemporal perspectives, but the motivation behind is usually being neglected, which is crucial to understanding activity changes. The objective of this study is to develop models and methods to better understand human activity patterns using crowdsourcing and geosocial media data. Thus, in this thesis, a method is first developed to detect activity changes, based on which a Markov chain-based model is developed to predict human movements. Then, semantics is introduced to uncover the motivation associated with the corresponding spatiotemporal patterns, which can infer what people do and discuss in a location at a specific time. Finally, human activity patterns are modelled from both spatiotemporal and semantic perspectives.

A 6-year GPS dataset of human movement in Beijing, China was used to evaluate the proposed predictive model. The results show that the predictive model can yield accurate prediction of the movement for those users who have significant activity changes (with R^2 improved from 0.295 to 0.762). A whole-year geo-tagged tweets posted within Toronto, Canada was acquired to analyze human activity patterns. A network model was finally created by the proposed approach to represent human activity patterns. The experimental findings demonstrate that most of the individuals (61%) have a regular activity pattern, while only a small number of people (10%) have a different activity pattern from the mass. With the inclusion of semantic information together with the spatiotemporal data as well as detecting the activity changes, such an approach can enhance the capability of human mobility and activity modelling, and thus pave the way for a more mechanistic understanding of how urban systems are being shaped, as well as how their sub-systems/components interact.

Key words: Human mobility, Human activity, Spatiotemporal pattern, Semantics, Markov chain, Topic modelling, Twitter

Acknowledgments

First and foremost, I would like to thank my supervisor Dr. Songnian Li for his guidance and support. I first met him in Central South University, Changsha, China where I received my master's degree before I came to Canada. I decided to work with him as his PhD student due to one of his research projects—event-driven GIS, which is an interesting, innovative and exciting project. However, during the preparing of my PhD candidacy exam, I felt that I was actually more interested in another topic (my PhD thesis now) which was not closely relevant to the project which we planned to work on. I talked to Dr. Li about the new topic and showed him some preliminary results, he accepted the new topic and fully supported me to pursue it during these four years without reservation. I deeply appreciate his physical and spiritual support particularly providing me a very free research environment. I also extend my appreciation to my other two supervisory committee members, Dr. Ahmed Shaker and Dr. Cherie Ding, for their constructive comments on both my PhD research proposal and this dissertation.

I am so indebted to have a number of professors in the department who gave me guidance throughout the study and shared their philosophy on research and life. I would like to thank Dr. Arnold Yuan for not only the free beers and meals often but also his attitude towards doing research and some impressive discussions for my first journal paper. I appreciate Dr. Joseph Y.J. Chow (Department of Civil & Urban Engineering, New York University), who was an Assistant Professor in Transportation Engineering in our department, for offering a part-time research work on a transportation project. This opportunity inspired me on how to integrate my work into transportation research. Dr. Michael Chapman was a senior professor whom I worked with as his Teaching Assistant for three years. I learned some useful life philosophy from our chatting. He was also one of my referees who supported me to apply to postdoc positions. I give him my sincere

thanks. I also appreciate Dr. Jingyuan Liu and Dr. Darko Joksimovic for sharing their interesting stories which made me more comfortable living in abroad.

I owe a number of thanks to the following individuals at Ryerson who lent me a hand through the study. I would like to thank Mr. Robin Luong since he was the kind of person who always made me relax and happy and he helped me a lot for easing into the community quickly when I first came to Toronto. I am thankful to Ms. Moh Moh Lin Khin who was a master's student at Ryerson University for her kind help when I first started here, and, certainly, the best home-made desserts in the world. Thanks also go to Mr. Desmond Rogan who always promptly fixed IT problems for our lab.

Next, I want to say something about Dr. Xintao Liu and Dr. Wai Yeung Yan who are two people with big impacts on my PhD study and my academic career. Xintao and I came to Ryerson University at the same time—September 2012. He was working with Dr. Songnian Li as a postdoc for two years after he completed his PhD in KTH, Sweden. I discussed with him a lot when I was preparing my research proposal. He gave me a lot of interesting and innovative ideas and constructive comments for my proposal without reservation. In fact, the original idea for my first journal paper was came up during our daily discussion and brainstorm, which laid the foundation to my subsequent studies and research work. He is also the first person other than my supervisor who taught me how to do research. He is going to the Hong Kong Polytechnic University as an assistant professor soon, I want to say, thank you for your support and encouragement all the way! I could not have done this thesis without you! I wish you all the best!

Dr. Wai Yeung Yan had just completed his PhD when I joined Ryerson University. I knew him due to his big achievements during his PhD study—Governor General's Academic Gold Medal winner and unbelievable publication and award records. Yan is the other person who boosted me to achieve a higher goal in terms of both academic career and life. He was more like a mentoring

teacher who taught me fundamental and necessary skills of doing research. More importantly, I learned from him that as long as you do everything with a positive purpose, you definitely will get what you want one day. Thank you for teaching me how to use Latex, proofreading my papers again and again every time, discussing my research ideas, taking care of me all the time and, most importantly, bringing me to all the best Cantonese restaurants around Toronto (although most of the time I actually drove you there). You are undoubtedly my role model.

The education from the schools that I attended created the knowledge foundation and the way of solving problems which formed the basis of this thesis. I would like to especially thank the teachers for their enthusiasm in teaching and ethical education from Beidajie Primary School (北大街小学) and Qinghai Huangchuan High School (青海湟川中学), Xining, China. I would love to thank Prof. Min Sun who was my master supervisor in Peking University, Beijing, China for his training in terms of the ability of solving problems and programing.

Lastly, I would like to thank my family for their support. I want to especially thank my father-in-law, Prof. Yongnian Zeng, who is somehow my supervisor for promptly updating me the state-of-the-art work related to my research area and the guidance of doing research. Also, I would like to thank my mother-in-law for taking care of my wife when I was away. I would like to thank my parents for providing an open and a loving environment in which I can still keep the sense of having interesting ideas popping up and focus on my study. I am very grateful to my wife, Dr. Yan Zeng, for encouraging me to pursue my dream and taking care of our parents for us. I love you forever!

Wei Huang

September 2nd, 2016

Dedication

This dissertation is gratefully dedicated to my late grandfather and maternal grandfather.

Table of Contents

Abstract.....	iii
Acknowledgments.....	v
Table of Contents	ix
List of Tables.....	xii
List of Figures.....	xiii
Chapter 1 Introduction.....	1
1.1 Research motivation.....	1
1.2 Preliminaries	3
1.3 Research objectives.....	3
1.4 Contributions.....	5
1.5 Overall methodology	7
1.6 Organization of this thesis	8
Chapter 2 Existing Work	10
2.1 Human mobility modelling	10
2.2 Human mobility prediction	13
2.3 Crowdsourcing data and geosocial media data	15
2.4 Discussion	17
Chapter 3 Human Mobility Prediction	19

3.1	Methodology	19
3.1.1	Data	20
3.1.2	Extraction of human activity	21
3.1.3	Activity clustering	23
3.1.4	Human activity change detection	26
3.1.5	Prediction of human movement	31
3.2	Results and discussion	34
3.2.1	Human activity analysis	34
3.2.2	Activity clusters exploration	38
3.2.3	Predictive results and evaluation	41
3.3	Summary	44
Chapter 4	Human Activity Motivation Modelling and Analysis	45
4.1	Methodology	45
4.1.1	Preliminary definitions	45
4.1.2	Probabilistic topic models	46
4.1.3	Spatiotemporal-semantic model for human activity patterns	47
4.2	Results and discussion	55
4.2.1	Data	56
4.2.2	Pattern inferring	57

4.2.3	Analysis of activity pattern similarity.....	61
4.2.4	Discussion.....	65
4.3	Summary.....	68
Chapter 5	Human Activity Pattern Modelling and Analysis.....	70
5.1	Methodology.....	70
5.1.1	Data.....	70
5.1.2	Spatiotemporal clustering.....	73
5.1.3	Semantic pattern modelling.....	74
5.1.4	Similarity among spatiotemporal activity clusters.....	74
5.2	Results and discussion.....	77
5.2.1	Spatiotemporal patterns.....	77
5.2.2	Semantic patterns.....	80
5.2.3	Activity patterns.....	83
5.3	Discussion.....	85
5.4	Summary.....	88
Chapter 6	Conclusions and Future Work.....	89
6.1	Conclusions.....	89
6.2	Future work.....	90
References		93

List of Tables

Table 1.1 The methodology of this thesis	7
Table 3.1 The number of single activities of each user	22
Table 3.2 Transition probability between activity clusters of user #153	43
Table 3.3 Transition probability between activity clusters of user #128.....	44
Table 4.1 The distribution over stayed places for sets of stayed places	59
Table 4.2 The distribution over words for topics.....	60
Table 4.3 An example of activity patterns of two users	61
Table 4.4 Similar spatiotemporal patterns between two users.....	66
Table 4.5 Two different semantic patterns associated with the spatiotemporal pattern in Table 4.4	67
Table 5.1 Semantic pattern of a selected spatiotemporal cluster as an example	82
Table 5.2 Top 10 word distributions in topic 56.....	82

List of Figures

Figure 3.1 The methodology flowchart	20
Figure 3.2 Trajectories of all users and a specific user #153 in Beijing, China	21
Figure 3.3 The geographic distribution of all users' single activities in Beijing, China	23
Figure 3.4 Spatiotemporal clusters	24
Figure 3.5 Distribution of time intervals between single activities	25
Figure 3.6 Similar activity clusters	29
Figure 3.7 An example Markov chain	31
Figure 3.8 Markov chains at different times	33
Figure 4.1 The topic model-based graphical models	50
Figure 4.2 Structure of human activity pattern	51
Figure 4.3 Toronto land-use category	57
Figure 4.4 Perplexity results on spatiotemporal pattern model and semantic pattern model	58
Figure 4.5 Activity pattern similarity.....	63
Figure 4.6 Distribution pattern of KL ' values	64
Figure 4.7 The number of user pairs in different KL ' value groups	65
Figure 5.1 Distribution of tweets' volume posted by each user and mapped tweets.....	72
Figure 5.2 Spatiotemporal activity clusters	73
Figure 5.4 Similar activity clusters	75
Figure 5.5 Distribution over tweet number in spatiotemporal clusters.....	78
Figure 5.6 Spatiotemporal cluster distribution.....	79
Figure 5.7 Activity temporal pattern in different spatial clusters	80

Figure 5.8 Perplexity distribution over topic number	81
Figure 5.9 Activity patterns depicted as a network.....	83
Figure 5.10 Distribution of number of spatiotemporal clusters among KL'	85
Figure 5.11 Connectivity distribution over population.....	87

Chapter 1 Introduction

1.1 Research motivation

Human activities influence how urban systems formulate. Urban systems, on the other hand, mediate feedback to human activities through a diversified set of processes and mechanisms, including movements, behaviors, thoughts, languages, psychology, etc. Predicting where individuals travel and modelling how people live and play in cities thus play an essential role in various urban applications, such as transportation planning and traffic forecasting, urban planning, public health and safety, emergency response, and so on. Scholars from related fields, such as transportation, geography, urban planning, computer science, geographical information science (GIS), social science and epidemiology, have recently placed intensive efforts on addressing the aforementioned two issues: 1) predicting where individuals travel and 2) modelling how people live and play in cities (Hanson, 2005; Merler & Ajelli, 2010; Song et al., 2010; Kang et al., 2012; Bogomolov et al., 2014; Ferrara et al., 2014; Alexander et al., 2015; Fast et al., 2015; Huang et al., 2015).

Two key components, space and time, are usually used to understand human activity patterns, which aim to model human mobility patterns in order to infer and predict where individuals intend to move. With the inferred and predicted human movements, urban systems can be better designed and evolved accordingly and some negative social incidents happening can be sensed and avoided. For example, intelligent transportation systems can be designed and configured in a way to control traffic lights and dispatch buses based on the trends of human movements. The area that would be affected by a certain type of infectious disease can be predicted based on the mobility pattern of people being involved. In order to detect human mobility patterns, relevant data, such as survey

and interview data of daily travel records (Kwan, 1999), global positioning system (GPS) trajectories of human movement (Zheng et al., 2008), and phone call data (Alexander et al., 2015), have been recently explored. These data have represented a boon to researchers, giving them a tool to uncover human activity patterns from spatiotemporal dimensions.

However, what is more important is why people move to or stop at a certain place. Is the motivation behind similar mobility patterns of different people the same? Especially with the emerging of multifunctional buildings in cities in recent decades, a variety of activities can be attended in the same location. For instance, individuals are allowed to enjoy leisure, work, and hang out in the same building.

Recently, massive multi-dimensional data are being unprecedentedly generated by various smart devices (e.g., smart phones and wearable devices) through APPs installed (e.g., social media and health APPs) (Li et al., 2015). People are more intended to share instantaneously what's happening and what they are interested in through texting, photographing or other ways through social media tools (e.g., Twitter, Facebook, and Instagram). Among which, texting is one of the favorite approaches to share such information (e.g., Twitter allows user to post no more than 140 words). Thus a variety of topics about activities, events and incidents happening summarized from the text may reflect what people do, think and discuss (Pak & Paroubek et al., 2010; Sakaki et al., 2010; Becker et al., 2011; Bertrand et al., 2013). These intensive and informative fine scale data associated with spatiotemporal and semantic information open many new doors to better understanding the motivations underlying human mobility patterns (Birenboim & Shoval, 2015). If the motivations laid behind can be modelled, similar mobility patterns can be further distinguished from each other, which may provide a new way to better understand how urban citizens live, work and play in cities.

1.2 Preliminaries

- **Human mobility and human movement:** where people travel over time. Two dimensions, time and space, are involved. Human mobility and human movement are usually alternatively used by scholars, so in this thesis they are equally used.
- **Human activity:** what people do, think and discuss in a location at a time. Semantics is used to represent what people do, think and discuss. Thus, three dimensions, time, space and semantics, are involved.

Human activity can be considered as human mobility together with the associated semantics in this thesis. Spatiotemporal patterns of human activities are equally used with human mobility patterns in this thesis.

- **Geosocial media data:** the social media data tagged with geographical coordinates.
- **Semantics:** topics inferred from textual information generated from certain places. A detailed definition can be found in Chapter 4.1.1

1.3 Research objectives

The overall objective is to develop models and methods to better understand human activity patterns using crowdsourcing and geosocial media data, where human movement can be precisely predicted and human activity patterns can be modelled in a better way. Three specific objectives aimed to be achieved are listed as follows:

- 1) **To develop a predictive model of human movement with capability of dealing with human activity changes.**

One of the challenges limiting the predictive accuracy is that a location of an activity is likely being replaced by another location due to certain reasons, such as home move, school transfer or job change. Such location changes of a type of activity are being treated as failure factors in training samples prior to the prediction. During the training process, these failure factors may still be given a high transition probability although they should be actually excluded from the prediction, resulting in a negative impact on movement prediction. To tackle this problem, a method that can detect these failure factors beforehand needs to be developed to aid in the human movement prediction.

In fact, these failure factors can only be detected if they exist in the training samples. Those failure factors, happening beyond the training samples, would still cause a negative impact on the prediction. To better deal with the failure factors and understand human mobility patterns, uncovering the motivations behind human mobility patterns become essential. The mobility patterns of individuals are usually considered as their spatiotemporal attributes, the information illustrating what they do, think and discuss can be regarded as semantic attributes representing the motivations behind human mobility patterns. These semantic attributes are the motivations forming human mobility patterns, which cannot be neglected for illustrating human mobility and activity patterns. As a result, it leads to the following voids that have to be investigated:

2) To model motivations behind human mobility patterns (spatiotemporal patterns of human activities) based on semantic information.

What people do and discuss in a specific location at a time can be represented by semantic information, thus the motivations underlying human mobility patterns can be inferred by the associated semantic patterns. In addition, a hypothesis—similar human mobility patterns may have difference motivations—aims to be quantitatively proven, which can help stress the importance of

the role of the motivations in modelling human activity patterns. Subsequently, the other objective is:

- 3) To model and uncover human activity patterns based on not only spatiotemporal dimensions but also semantic dimension.**

Based on Objective (2), human activity patterns are then modelled and analyzed with emphasis on the semantic patterns, representing the motivations, associated with the corresponding spatiotemporal patterns, which can illustrate where and when people stay and what they do, think and discuss.

1.4 Contributions

Through achieving the above objectives, this thesis has the following four major contributions:

- 1) A new approach for detecting human activity changes.** Activity changes play a negative role in human movement prediction. If such activity changes can be detected, the predictive accuracy can be significantly improved for individuals who are involved in activity changes. The developed approach has a capability of detecting two types of activity changes, where most of existing researches did not consider such a factor in human movement prediction.
- 2) A human movement predictive model based on Markov chain with consideration of human activity changes.** Most of the existing human movement prediction models did not consider time component for prediction, where the location component was mainly addressed. The developed predictive model has a capability of predicting human movements according to not only current location but also current time. In

addition, the model can successfully avoid the negative impacts of human activity changes on predictive accuracy with the approach being introduced in Contribution (1), leading to a significant improvement on predictive accuracy for the human movements having human activity changes.

- 3) **A new method to incorporate semantic information associated with human mobility patterns to infer the associated motivation.** Majority of existing work on modelling human activity patterns mainly focused on spatiotemporal patterns, but the associated motivations were usually being neglected. Through uncovering the motivations behind the associated spatiotemporal patterns of human activities, the human activity patterns can be better understood. To fill this gap, semantic information representing what people do and discuss in a location at a time was used to infer the motivation. Through incorporating motivations into the associated spatiotemporal patterns, it can quantitatively prove that there exist certain similar human mobility patterns that are shaped by different motivations. Based on that, similar human mobility patterns can be further distinguished.
- 4) **A new approach to model human activity patterns from space, time and semantics.** Based on Contribution (3), human activities can be analyzed from space, time and semantics dimensions. The similarity in terms of time and semantics between human activities can be computed, based on which a network consisting of human activities can be designed to represent human activity patterns. The network provides a new way to better understand how people live, work and play in cities.

1.5 Overall methodology

Table 1.1 shows an overview of the research methodology used in this thesis. Through combining the review of the state-of-the-art journal papers and conference papers in the past decade, the research challenges regarding human movement predictions are identified and then research Objective (1) is proposed. To achieve research Objective (1), a new workflow of human movement prediction is proposed and evaluated with GPS trajectories collected for human movements with a duration of six years. More specifically, spatiotemporal clustering and analysis are implemented for data processing and human activity change detection, then a Markov chain model is used to develop the human movement predictive model based on the human movement GPS trajectories.

In order to deal with the human activity changes happening out of training samples, such as predicting the changes instead of detecting the changes, as well as better understanding how people live in cities, research Objective (2) and (3) are subsequently investigated. Machine learning algorithm is used to model the spatiotemporal patterns of human activities and the associated semantic patterns in Objective (2). The hypothesis in Objective (2) is quantitatively justified based on a case study on analyzing geo-tagged tweets collected within Toronto, Canada.

Through building on top of Objective (2), spatiotemporal clustering and machine learning are used to further analyze human activity patterns based on a whole year geo-tagged tweets in Toronto. Subsequently, Objective (3) is achieved by computing and analyzing the similarity in terms of time and semantics between human activity clusters.

Table 1.1 The methodology of this thesis

Research Objectives	Methodology
---------------------	-------------

1) To develop a predictive model of human movement with capability of dealing with human activity changes.	Spatiotemporal clustering, spatiotemporal analysis and Markov model
2) To model motivations behind human mobility patterns (spatiotemporal patterns of human activities) based on semantic information.	Machine learning and statistical analysis
3) To develop a method to uncover human activity patterns based on not only spatiotemporal dimensions but also semantic dimension.	Spatiotemporal clustering, machine learning and statistical analysis

1.6 Organization of this thesis

The thesis contains six chapters. Following the introduction in Chapter 1, Chapter 2 presents existing work of the state-of-the-art development in human mobility modelling as well as human mobility predictions, and identifies the current research voids so as to pave the way for the following work presented in the subsequent chapters.

Chapter 3 presents an approach of human activity change detection, based on which a Markov chain-based model is developed to predict human movements. A dataset of 6-year human GPS trajectories is used to test the accuracy of the proposed predictive model.

Chapter 4 presents a method to model human mobility patterns (spatiotemporal patterns of human activities) and the associated semantic patterns representing the motivation. The hypothesis—similar human mobility patterns may have different motivations—is proven by analyzing the

relationship between the spatiotemporal patterns and spatiotemporal-semantic patterns of human activities based on a geo-tagged tweets data of Toronto.

Chapter 5 represents a model to further analyze human activity patterns from space, time and semantics based on Chapter 4. Some findings are discussed based on analyzing a whole-year geo-tagged tweets in Toronto. Conclusions of the thesis are drawn in Chapter 6, alongside with directions for future work.

Chapter 2 Existing Work

Quantitative understanding and modelling of human mobility patterns are the drivers of a variety of urban applications. Ratti et al. (2006) pointed out that analyzing human activity patterns lays a foundation in urban planning. Some other urban-related researches, such as transportation and epidemic, also addressed the importance of a better understanding of human mobility patterns. For example, Pan et al. (2013) proposed to identify traffic anomalies by detecting the drivers' routing behavior patterns which are significantly different from their original patterns. Work et al. (2008) discussed the impacts of human mobility on the automotive Cyber Physical System (CPS). They claimed that new human mobility models need to be developed in order to enable the automotive CPS to be more open and flexible in the large context of human mobility. In addition, Belik et al. (2011) developed a model for spatial epidemics with consideration of human mobility patterns. Frias et al. (2011) proposed an agent-based system based on individual mobility patterns to model virus spreading. Poletto et al. (2013) analyzed epidemic processes based on where individuals travelled and stayed. Apparently, the aforementioned studies take human mobility patterns into account, thus modelling human mobility and understanding the mobility patterns have been studied by scholars from different fields.

2.1 Human mobility modelling

The study of human mobility has a long history. To our knowledge, the earliest published research on understanding human mobility dates back to the eighteenth century—The Laws of Migration (Ravenstein, 1885). Census data was used to explore the migration. Due to the limitation of the way to obtain human mobility data at early stage, coarse-scale dataset was largely used for

exploring human mobility patterns (Kwan, 1999). Some classic models to understand how people migrate were developed. For example, some of researchers believe that human migration follows Newton's law of gravity, where the mobility of individuals decreases with the physical distance between two locations (Carrothers, 1969; Erlander & Stewart, 1990; Sen & Smith; 2012). According to that, a model called "gravity model" was developed to predict human mobility. In contrary, some other studies argued that the travel distance from origin to destination depends on the needs giving rise to its journey rather than the physical distance between the origin and the destination, which calls "Intervening opportunities" (Stouffer, 1940). In addition, Brockmann et al. (2006) estimated individual displacement by analyzing circulations of bank notes in U.S.A. and proposed a scaling law for modelling human mobility patterns. They addressed that the jumping size and waiting time in user-generated trajectories follow a power law distribution, where long travel distances and stay time are rare while short ones are common, such a phenomenon follows the so called "Lévy-flight" patterns. Such patterns are claimed to be consistent at any reasonable spatiotemporal scale.

With the emerging of the pervasive computing last decade, human traces have been recorded in a high-resolution manner, such as those collected by GPS devices (Zheng et al., 2008), mobile phone calls (Alexander et al., 2015) and geo-tagged social media (Li et al., 2015), which provide an opportunity to better understand the way people move in a quantitative way. Gonzalez et al. (2008) used phone calls or text messages and the location of the signal towers to reconstruct users' time-resolved trajectories. Their research found that people are intended to return to a few most frequently visited locations (Lévy flight), which can be characterized by a single spatial probability distribution that indicates the dynamics behind the reproducible scaling patterns. Candia et al. (2008) discovered that the time interval of consecutive phone calls followed a heavy-tailed

distribution. Later on, Song et al. (2010a, 2010b) further explored the regularity of human travels using the same data and proposed two generic mechanisms, i.e., occasional exploration and preferential return that govern human trajectories from a microscopic perspective, and they further claimed that 93% of potential predictability in human mobility. Simini et al. (2012) proposed a universal model—radiation model—for mobility and migration patterns. Comparing to gravity model, the radiation model is parameter free that can be applied in areas lacking of previous mobility measurements. Schneider et al. (2013) applied graphs to investigate the underlying mechanisms of daily human mobility patterns. Directed networks were used to represent daily trajectories of individuals. Based on human mobility patterns, efforts have been made on comparing the similarity of human mobility (Yuan and Raubal, 2014, Xia et al., 2011, Eagle et al. 2009, Joh et al. 2002) since it may contribute to understanding of the social interaction among different demographic groups, and generating of specific recommendations in social network applications (Zheng et al., 2011).

Through analyzing human mobility patterns, human activities can be further uncovered and extracted (Zheng et al., 2009). For example, Ashbrook and Starner (2003) suggested that the most significant places where a user stays should be those the user spends some of his/her time (i.e., a user stays somewhere for a certain amount of time to do something). They extracted user activities with a given time threshold (e.g., 20 minutes). Similarly, Liao et al. (2006) extracted and clustered user activity locations and proposed sophisticated Markov models to recognize and infer human activities. Their method is capable of learning specific motion patterns of transportation routines to detect user errors. Based on the extracted activities, related applications, such as friends and location recommendation based on individual location history (Zheng et al., 2011), shop recommendation based on shopping location histories (Takeuchi & Sugimoto, 2006), and

restaurants recommendation based on users' preferences (Park et al., 2007), were developed. More importantly, human movement prediction can be better performed.

2.2 Human mobility prediction

According to the manner by which data are modelled, the previously-developed methods for predicting human movement can be classified into three general distinct approaches: state-space models, data mining techniques and template matching techniques. Furthermore, some other methods combining different methods together are viewed as hybrid methods.

State-space models attempt to capture the variation in spatial sequences through sequence models such as generative Hidden Markov Models (HMMs) (Rabiner, 1989), discriminative Conditional Random Fields (CRFs) (Vail et al., 2007; Sutton & McCallum, 2010), or extensions of these two well-known approaches (Dietterich, 2002). Generally, these models have been used successfully in dealing with uncertainty, but they also suffer from high level of training complexity. In the case of mobility prediction, state-space models such as HMMs can naturally be used, since they support the generation of possible future visits and the estimation of an associated probability. For example, Asahara et al. (2011) proposed a state-space modelling method for predicting pedestrian movement on the basis of a mixed Markov-chain model (MMM), taking into account a pedestrian's personality as an unobservable parameter, together with the pedestrian's previous status. The authors reported an accuracy of 74.4% for the HMM method. In a comparison over the same dataset, the authors reported that methods based on Markov-chain models, or based on Hidden Markov Models, achieve lower prediction rates of about 45% and 2%, respectively, for each of these two cases. Moreover, Gambs et al. (2012) extended a previously proposed mobility model called the Mobility Markov Chain (MMC), in order to consider the n previous visited

locations. This research essentially corresponds to a higher order Markov model. Experiments on different datasets showed that the accuracy of the predictions ranges between 70% and 95%, but they also show that improvements obtained by increasing $n > 2$ do not compensate for the computational overhead. Mathew et al. (2012) proposed a predictive model based on Hidden Markov model (HMM), which uses a timestamp (i.e., 7 a.m. to 7 p.m. of weekday, 7 p.m. to 7 a.m. of weekday, and weekend) to associate the places in sequences.

Data mining techniques, on the other hand, explore frequent patterns and association rules, by defining a trajectory as an ordered sequence of time-stamped activities, and using sequence analysis methods such as modified versions of the Apriori algorithm (Morzy, 2007; Monreale et al., 2009). Most existing data mining methods attempt to maximize confidence based on previous occurrences (i.e., they do not generalize as good as state-space models), and they often do not consider any notion of spatial and/or temporal distance. For example, Morzy (2007) proposed data mining methods for predicting the future location of moving objects. Association rules from the moving object database are extracted and, thus offer a previously unseen trajectory of a moving object. Matching functions were used to select the best association rule that matches the trajectory for subsequent prediction. An accuracy of 80% was reported with the best configuration of the proposed system.

The third type of approaches, which are based on template matching, compare extracted features to pre-stored patterns or templates, using similarity metrics specific to sequential or time-series data. These similarity metrics include dynamic time warping and other sequence alignment approaches that are essential variations of the edit distance computed between the sequences, e.g., edit distance with real penalty or edit distance on real sequence. They also include algorithms based on the longest common subsequence, or even other heuristic algorithms similar to those used

in more traditional streaming matching problems (Riedel et al., 2008; Ossama & Mokhtar, 2009). Template matching techniques have also been reported to have certain issues with high runtime complexity, noise intolerance, or in dealing with spatial activity variation.

Furthermore, hybrid sequence analysis approaches, combining multiple types of information, have been proposed. For instance, Jeung et al. (2008) proposed a hybrid prediction approach which estimates future activities based on pattern information extracted from similar trajectories, together with motion functions based on recent movements. Lu and Tseng (2009) used a sequence similarity measure to evaluate the similarity between activity histories, and used a clustering algorithm to form a user cluster model of the activity histories, based on the similarity measure. The clusters were then used to predict the movement of individuals, i.e., their next activity. Ying et al. (2011) proposed an approach for predicting the next activity of an individual based on both geographic and semantic features of trajectories. They used a cluster-based prediction strategy that evaluates the next activity based on the frequent behaviors of similar users in the same cluster, as determined by analyzing common behavior in terms of the semantic trajectories.

2.3 Crowdsourcing data and geosocial media data

Recently, ubiquitous computing has proven to be a convenient way that continually collects individual's activity data. Smart devices, such as smart phones and wearable GPS receivers, can record the location and time where people stay. For example, phone call data can be used to indicate the location and time when an activity happens. The coordinates of the location where users making a call can be calculated based on the location of towers routing the communication. This type of data seems to be popular for understanding the way people travel in cities (Gonzalez et al., 2008; Song et al., 2010a, 2010b; Alexander et al., 2015). On the other hand, assisted GPS

equipped with smart phones and wearable GPS receivers have provided a more precise way to collect individuals' daily movement trajectories. In 2007, Microsoft Asia launched in the GeoLife project which enables users to share their GPS trajectories of movement through smartphones and wearable GPS receivers (Zheng et al., 2008). Through analyzing the massive trajectories shared by users, human daily activities and mobility can be extracted and understood (Zheng et al., 2008, 2009). These crowdsourcing datasets have provided a new opportunity for better understanding human mobility and activity patterns.

Social media, like Twitter, Facebook and Instagram, enables users to instantly share what they are doing, thinking and discussing at a certain place and time by means of texting or photographing. More importantly, the current location where users are posting is allowed to be shared as coordinates (Longitude and Latitude), leading to a new way to understand the way people live, work and play in cities. Thus, those social media tagged with geo-coordinates, called geosocial media, has been proposed as proxies for providing valuable information on human mobility (Austin et al., 2014; Hawelka et al., 2014; Wu et al., 2014). For instance, more than 500 million users from Twitter post over 340 million tweets per day, where current location can be tagged with tweets. Therefore, geotagged tweets have been considered as one of ideal data sources to study human mobility, although they are still in debated about their quality and representation. In addition to geo-location and time, there exists one more dimension, i.e., textual information or imagery, embedded in the geosocial media data, which opens a new door for better understanding human mobility as well as the mechanism forming certain mobility and activity patterns.

2.4 Discussion

The aforementioned studies on modelling human mobility patterns can only discover some activity patterns of individuals from spatiotemporal aspects. In fact, what people do and discuss at a place and a time cannot be neglected since activity patterns and human movements are cross-correlated. If a mobility pattern is illustrated as a type of expression of human behavior in terms of spatiotemporal dimensions, the mechanism or motivation that leads to the formation of such expression should be reflected by the specific semantic information (i.e., texts recording normal life and opinion on current issues and events). If the motivation can be inferred or uncovered by any means, how people move can be more reasonably and accurately predicted that may contribute to better design of our urban systems. To fill this gap, this thesis attempts to determine the motivation behind human mobility patterns by analyzing the associated semantic patterns, and then model human activity patterns from space, time and semantics to analyze how people live, work and play in cities.

On the other hand, the studies discussed in the previous section either ignored or just simply used time dimension to predict human movement. In fact, time dimension should be considered to detect human activity changes because some historical human activity changes, such as home change, school transfer or job change, have negative impacts on movement prediction. For example, Song et al. (2006) introduced an aging mechanism to reduce the negative impact of changed locations toward improving the prediction performance. A multiplicative factor was proposed to intervene in the process of sample training, and to reduce the contribution of old samples where changed locations are involved. Etter et al. (2013) proposed an algorithm that is able to detect home changes. In these two studies, the datasets with semantic information were used to detect home changes. The authors assumed that each person has a single home, and once another home is

subsequently detected during the model training, the detected home is flagged as the new home. Although these studies investigated the historical changed locations (actually they attempted to detect home changes only), they can only deal with datasets with semantic information, which is not applicable for most user-generated trajectory datasets. To fill the voids discussed above, an algorithm of human activity change detection is developed, based on which a human movement predictive model is developed with consideration of the activity changes.

Chapter 3 Human Mobility Prediction

This chapter presents a methodology of human movement prediction from spatiotemporal perspectives by taking into account the impact of activity changes. The changes of human activity are first analyzed and defined, then a method to detect such changes is proposed, based on which a Markov chain model is used to predict human movement. The Microsoft GeoLife dataset was used to test the methodology, and the data of two selected users were used to evaluate the performance of the prediction.

3.1 Methodology

A four-step methodology is proposed (Figure 3.1), which consists of: 1) extracting human activities from raw GPS trajectories, 2) clustering activities in terms of spatiotemporal dimensions using density-based spatial clustering of applications with noise (DBSCAN) method (Ester et al., 1996), 3) detecting two types of human activity changes based on the identified activity clusters, and 4) predicting human movement based on Markov chain model with consideration of activity changes.

Figure 3.1 The methodology flowchart

3.1.1 Data

The dataset, obtained from the Microsoft GeoLife project, contains 17,621 GPS trajectories that were collected by 182 users over 6 years (from April 2007 to August 2012) using different wearable GPS receivers and cell phones (source: <http://research.microsoft.com/en-us/projects/geolife/>). These trajectories contain 25,298,541 GPS points, covering a total distance of about 1.2 million kilometers and a total duration of more than 48,000 hours (c.f., Zheng et al. 2009 for more details). These trajectories recorded a broad range of users' outdoor movements, not only daily routines such as staying at home and working, but also leisure activities such as shopping, sightseeing, dining, hiking and cycling. Figure 3.2 illustrates the trajectories of all users as well as a specific user #153 in Beijing, China (lines in color). The color shows the density of the GPS points determined based on a 200 m by 200 m window.

Figure 3.2 Trajectories of all users (a) and a specific user #153 (b) in Beijing, China

3.1.2 Extraction of human activity

A human activity can be defined as a meaningful location of a user where he/she spends some time (Ashbrook and Starner, 2003). The method proposed by Zheng et al. (2009, 2011) is used to extract human activities from raw GPS trajectories based on a time threshold of 20 minutes and a distance threshold of 200 meters. This means that an activity is defined if a user stays for at least 20 minutes in a geographic region with a diameter of 200 meters.

Definition 3.1. A **single activity** is an activity (point) extracted from raw GPS trajectories.

In total, 38,520 single activities are extracted from 182 users over 6 years (see Table 3.1). Some users do not have any activities, while 7 users have more than 1,000 activities. Single activities of all users are plotted in Figure 3.3. According to Figure 3.3, it is obvious that most of single activities are located in the north-west part of Beijing, China, where a few universities and research

institutes (e.g., Peking University, Microsoft Research Asia, etc.) are situated. Therefore, it can infer that most of users come from these academic institutions.

Table 3.1 The number of single activities of each user extracted from the Microsoft GeoLife dataset

User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount	User ID	Activity Amount
153	2951	140	424	20	176	134	100	95	48	154	25	76	12	27	4
128	2026	52	371	15	170	155	98	94	48	114	24	54	12	148	3
4	1487	92	349	12	164	40	93	66	48	93	24	156	11	72	3
3	1233	104	341	179	151	74	92	57	48	79	24	127	11	149	2
17	1122	34	323	13	151	147	85	23	48	176	21	77	11	137	2
163	1113	37	319	9	151	83	78	141	47	106	21	80	10	116	2
68	1050	112	296	71	148	131	75	55	44	164	19	143	9	31	2
30	993	36	292	43	141	33	73	102	42	145	19	100	9	21	2
167	783	42	290	73	138	29	66	169	41	99	19	124	8	177	1
85	751	24	251	174	135	122	65	56	40	59	18	121	8	172	1
144	680	5	249	125	132	81	65	181	38	161	16	63	8	133	1
41	663	67	232	10	129	130	64	58	38	98	16	45	8	60	1
22	610	96	225	91	128	89	64	46	37	61	16	117	7	48	0
25	607	65	214	28	126	6	63	138	36	129	15	87	7	49	0
39	591	115	210	78	124	159	62	157	34	158	14	70	7	66	0
0	562	82	195	16	124	103	59	113	33	146	14	53	7	118	0
62	548	44	186	168	123	165	57	105	31	108	14	175	6	120	0
14	511	11	186	18	117	32	57	135	28	47	14	166	6	123	0
35	503	1	185	19	112	64	56	111	28	152	13	162	6	132	0
126	490	7	184	51	109	97	53	139	27	90	13	109	6	160	0
38	451	142	182	26	107	88	53	136	26	173	12	170	4	180	0
84	434	119	179	8	106	110	50	75	26	171	12	151	4		
2	433	50	177	20	176	150	48	69	26	86	12	107	4

Figure 3.3 The geographic distribution of all users' single activities in Beijing, China

3.1.3 Activity clustering

Normally, GPS receivers are not able to record the identical coordinates at the same location due to their positioning error, even though users precisely stay at the same point of a place. More importantly, a series of activities may occur at different times, i.e., their arrival times are different (e.g., the time of arriving home can be either in the noon or in the evening). Therefore, it is necessary to perform spatiotemporal clustering to classify single activities considering temporal aspect and further filter single activities in terms of spatial dimension, which is the key to detect human activity changes and build a predictive model in our work later on.

Definition 3.2. An **activity cluster** is a group of single activities happened at the same location in a specific time duration.

Specific time and space thresholds are used to define search radius and set 1 as the minimum number of points required to form a dense region. The DBSCAN is applied to perform the required clustering analysis. If the “distance” in terms of time and space amongst points is within the specified thresholds, these points are grouped as a single cluster (Figure 3.4). The accuracy of GPS receivers is used as the space threshold since the deviation between two coordinates of the same location is mainly caused by the error of GPS receivers. A threshold of 15 meters, therefore, is set as the space threshold in this paper, since it is the average error of the GPS receivers used in the GeoLife project. Since there is no universal number that could be used to describe each user’s time characteristics, a data-driven method is developed to determine this threshold for different users.

Figure 3.4 Spatiotemporal clusters

The distribution of time interval between a single activity and the subsequent single activity varies with different individuals but overall follows a Heavy-tailed distribution (Figure 3.5). According to Jiang et al. (2013), if a series of variables follow a Heavy-tailed distribution, the mean can separate all the variables into a high percentage of small ones and a low percentage of large ones. Thus, the mean can represent the characteristics of the major population. Based on this theory, the mean of time intervals (between arrival time and leaving time at a place) carries some underlying meanings: similar and continuous single activities happen within a certain period of time. In other words, single activities happened around a time within the mean of time intervals have similar characteristics in terms of time aspect. As a result, the mean of time intervals is used as the time threshold to cluster the corresponding user's single activities. The two thresholds are thus identified: 15 meters (as mentioned above) as the space threshold and the mean of time intervals as the time threshold. Every single user may have different time threshold, which entirely depends on the user's temporal pattern of the extracted single activities.

Figure 3.5 Distribution of time intervals between single activities

If a single activity of one activity cluster and a single activity of another activity cluster are within the space threshold (15 meters), these activity clusters are considered as located in the same place, which is represented by the centroid of the single activities from these different activity clusters.

3.1.4 Human activity change detection

Human activity changes play a negative role in predicting human movement because a new location of an activity cluster recently emerged in the training sample has lower transition probability than the previous location within a short period. In other words, the previous location of that specific activity cluster still has a higher transition probability than that of the new location if the change occurred close to the end of the sample data for training predictive model. For instance, a person recently moves to a new apartment A_1 from the previous apartment A_2 . If the next possible place where this person is heading to is her/his home, the most possible location of her/his home is still the location of A_2 since it has a higher transition probability than the location of the apartment A_1 . Actually, the correct predicted location of the home should be the location of the apartment A_1 regardless how high the transition probability of the apartment A_2 is. Hence, it aims to model such activity change by providing a definition of activity change and a change detection method.

Definition 3.3. A human activity change occurs when i) **activity cluster location changed**; or ii) **activity cluster obsoleted**. Some human activity changes, such as job change (the location of office changed) and school transfer (the location of campus changed), are related to the change of activity cluster location. Other human activity changes, such as graduation (the activity of studying ended) and retirement (the activity of working terminated), can be seen as activity cluster obsoleted.

In order to detect how activity cluster location changes, the two methods proposed by Song et al. (2006) and Etter et al. (2013) both can detect a person's home change through analyzing if the location of a home coming after the previous one in the training samples is the same as the location of the previous home. If the answer is no, the home of this person is changed. This kind of method is based on an assumption that every person has only one single home. When two homes are found in different locations, a home activity change is identified. Based on that, an assumption can be made that one person can only have one type of activity during specific time duration. For example, one person generally only has one single working place, school, or gym in a given time duration. Furthermore, since there is no additional information in GPS dataset indicating the type of place visited (e.g., home, office or school), it is impossible to directly obtain the name of the activities extracted from GPS trajectories. As a result, similarity between activity clusters is calculated and if similar activity clusters occurred sequentially in terms of date is examined. If they are similar and occur sequentially, they are considered as changed activity clusters. The advantage of the proposed method is that it can detect the changes of certain activities that are difficult to infer their semantics. Two parameters, arrival time of activity cluster and stay duration of activity cluster, are used to define the similarity of activity clusters (Figure 3.6a). If any two activity clusters occur at similar time and their stay durations do not have significant difference, such activity clusters can be considered as similar, representing a certain type of activity. To quantitatively define the similarity among activity clusters based on the work of Chen et al. (2011), the overlapping area between two rectangles enclosed by arrival time and stay duration of any two activity clusters is used to measure the similarity. More specifically, the equation proposed by Tversky (1997) based on a similarity theory in psychology is utilized to quantitatively define the similarity between two activity clusters, which can be expressed by:

$$D = \frac{c}{a + b - c} \quad (3.1)$$

where a and b are the area of the two rectangles enclosed by arrival time and stay duration, respectively (Figure 3.6b); and c is the overlapping area of a and b . If two activity clusters have same arrival time and stay duration (i.e., $a = b$), the similarity measure D equals to 1. In other words, the higher level the similarity between two activity clusters, the closer the value of D is to 1.

Figure 3.6 Similar activity clusters (Activity cluster a is similar with Activity cluster b). The arrival time of these activities is $Time1$, and the duration the person stays on activity a and activity b is $Time2 - Time1$

I (a); similarity measure (b); and obsolete activity cluster(c)

In this study, two activity clusters can be considered as similar activity clusters if $D > d$, where d is a threshold and can be computed by:

$$d = \frac{c}{\sum_{i=1}^n a_i - (n-1)c} \quad (3.2)$$

where a_i refers to the area of the rectangle enclosed by arrival time and stay duration of a single activity in an activity cluster; n refers to the number of single activities in that activity cluster; c denotes the overlapping area of all the activities in that activity cluster. In this case, d represents the similarity of all activities in the clusters, thus it can be used to determine if another activity cluster have the same similarity.

On the other hand, to detect *obsolete activity clusters*, the average time interval of single activities happening in an activity cluster is compared with the time difference between the last time a single activity happened in this activity cluster and the end time of training sample (see Figure 3.6c). The average time interval of single activities in an activity cluster can be defined as follows:

$$T_{int} = \frac{\sum Ta}{n-1} \quad (3.3)$$

where T_a refers to the time interval between two neighboring single activities in terms of happening time in an activity cluster; n denotes the number of single activities in an activity cluster.

The time difference can be defined as follows:

$$D_t = T_{end} - V_t(L) \quad (3.4)$$

where T_{end} refers to the end time of the training sample, $V_t(L)$ is the time of the single activity in the activity cluster happened last time L . If $D_t > T_{int}$, then this type of activities is considered as obsolete activity cluster, which means this activity cluster will not be taken into account for prediction. For instance, a dataset from January to June is used to train the predictive model. If single activities are found in one activity cluster occurred every day, i.e., the average time interval is about one day, but did not continue within the last month before the end of June, this activity cluster should be considered as obsolete and should not be used in prediction.

Based on the above discussion, an algorithm to detect the activity change is proposed. The pseudo-code of the process is shown in Algorithm 1.

Algorithm 1 Activity change detection algorithm

Input: *all activity clusters of a single user*

Output: *updated activity clusters of that single user*

begin

 detect similar activity cluster(s) using Eqs.(1) & (2) for each activity cluster;

If *there is no similar activity cluster found for some activity clusters* **then**

 calculate T_{int} using Equation (3.3) and D_t using Equation (3.4) for these activity clusters;

 compare D_t and T_{int} for these activity clusters;

 remove those activity clusters whose $D_t > T_{int}$;

end

 use the location of the newest activity cluster as the location of all its similar activity clusters;

 update the activity clusters

end

3.1.5 Prediction of human movement

Markov chain is a well-known model for analyzing sequential data, which can be defined as a sequence of random variables $X_1, X_2, X_3, \dots, X_n$ with the Markov property and, given the present state, the future and past states are independent (Markov, 1971). Formally, a Markov chain can be described as: $P_r(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P_r(X_{n+1} = x_{n+1} \mid X_n = x_n)$, where the possible values of X_i form a countable set S is called the state space of the chain. Markov chain is often described by a directed graph, of which the edges are labeled by the probability of going from one state to another state. Figure 3.7 illustrates an example of a classic Markov chain with three states describing by a directed graph. From X_1 to X_2 , X_2 to X_3 , X_3 to X_1 , the probability is p_1 , p_2 , and p_3 , respectively.

Figure 3.7 An example Markov chain

It aims to predict the human movement at a specific time and a specific place, which depends purely on the current location and time without considering the preceding states. It inherits the typical Markov property; therefore, a human movement prediction model is developed based on the concept of Markov chain.

First, all activity clusters are preprocessed by activity change detection algorithm (Algorithm 1), which enables changed activity clusters to be preprocessed. These preprocessed activity clusters

are then used to train Markov chains. The activity clusters are utilized to set up the states of Markov chains. The transition probability between two states denotes the probability of a specific user traveled from one place (the centroid of an activity cluster) to another at a specific time. This transition probability from state I to state J at time t is computed by:

$$P^t(J|I) = \sum T_{JI} / \sum_{k=m}^j \sum T_{ki} \quad (3.5)$$

where T_{JI} denotes a transition from state I to state J ; m refers to the state connected to state I . Each activity cluster is associated with a specific Markov chain. In this case, the Markov chains related to the same place may be different if more than one cluster exists at the same location. For example, in Figure 3.8, at location C , three activity clusters at different time, Time a , b and c , are found. Thus, there exist three Markov chains associated to location C . At Time a , the transition probability from location C to B and D are 90% and 10%, respectively, and 10% and 90%, respectively at Time c . At Time b , there is a probability of 80% traveling to location A , but no transitions to location A exist at the other two times.

Figure 3.8 Markov chains at different times

As a result, the location most likely to be visited from one location varies at different times. Given an activity cluster set $S = (s_1, s_2, s_3, \dots, s_n)$, the location most likely to be visited from location M can be defined as:

$$L(t) = \begin{cases} \operatorname{argmax}(P^{t_1}(J|M)) \\ \operatorname{argmax}(P^{t_2}(J|M)) \\ \dots \\ \operatorname{argmax}(P^{t_n}(J|M)) \end{cases} \quad (3.6)$$

Where $J \in S$, t_n denotes a specific time when some transitions are generated from location M .

Notice that the time t_n may actually be a very short time period since the time of each single activity occurred in one activity cluster may have difference in minutes. For instance, the time arrived home may be 10 to 15 minutes before or after 8:00 p.m. In this study, this very short time period is treated as a time point to simplify the description of the predictive model.

3.2 Results and discussion

In this section, the extracted single activities are first visually explored using ArcScene. The temporal patterns between two types of time attributes: arrival time and occurring duration of single activities are analyzed. Based on that, two top users (#153 and #128) are selected for experimental testing to evaluate the performance of the predictive model. The selection of these two users is mainly because they have two different typical mobility patterns. One user has obvious activity change pattern, while the other one has three identical activities without changes. Furthermore, the spatiotemporal distribution of their single activities is more diverse than other users.

3.2.1 Human activity analysis

To capture the spatiotemporal regularities of human activities towards an understanding of human mobility pattern, all users' single activities and top users whose single activities are over a thousand are visualized in space and time dimensions (Figure 3.9). The height of the points in Figure 3.9 refers to time of day when the activity happened in minutes in each year, with the red representing the most recent year 2012, and the blue representing the earliest year 2007. Therefore, the total height in Figure 3.9 is minutes in 24 hours multiplied by the number of years. The horizontal distribution of the points is based on longitude and latitude coordinates, which are plotted on top of the major roads in the Beijing (shown as grey lines). To better explore the movement patterns

and activity changes, a location is defined as a frequently and continuously revisited place as presented in a shape of vertical line in Figure 3.9. It is easy to note the obvious activity locations such as A and B for all users, A and B for user #153, and A, B and C for user #128. It should be noted that such locations of single activities found for both all users and individual users are aligned with the concept of spatiotemporal clusters.

Figure 3.9 Spatiotemporal distribution of single activities of all users and top users with more than 1,000 single activities: (a) All users; (b) User #153; (c) User #128; (d) User #4; (e) User #3; (f) User #17; (g) User #163; (h) User #68

For the single activities of all users, the geographic region of the single activities of all users (Figure 3.9a) in 2008 and 2009 is the most widespread and has higher density than any other years. Such a phenomenon may be mainly due to the 2008 Summer Olympics Games held in Beijing, China; the Olympic venues distributed around Beijing were visited more frequently in 2008 (the Olympics year) and 2009 (more tourism for the Olympics venues). Single activities for all users at location A and B just occurred for a short period during the year 2009. Such single activities are from different users, indicating some users temporally appeared in such locations (e.g., specific Olympics venues), and then they did not revisit again.

Those single activities associated with individual users (e.g., A and B for user #153 in Figure 3.9b) could carry some semantic information (such as home and working place). Nevertheless, these fixed locations could change over time. As shown in Figure 3.9b, the activity pattern of user #153 demonstrates a pattern of activity change according to the Definition 3.3, i.e., the activities changed from location A to location B in 2009. In contrast, the activity pattern of user #128 (Figure 3.9c) demonstrates a type of stable pattern, where he/she mostly stayed at location A, B and C. In addition, there are some similarities of human activity pattern if comparing the single activity distribution of different users. For example, user #4 has a similar activity pattern with user #3, as shown in Figure 3.9d and 3.9e, since their single activity distributions in terms of temporal dimension and spatial dimension do not show significant visual differences.

The temporal patterns are also plotted based on stay duration versus arrival time for the single activities of all users and top users who have over a thousand single activities in Figure 3.10. The vertical axis represents the stay duration of single activities in hours and the horizontal axis refers to the arrival time of single activities in Beijing time. Overall, all users have two significant stay durations of approximate 9 to 14 hours in total. One is related to the arrival time from around 8:00

a.m. to 9:30 a.m., marked by the black circle in Figure 3.10 (All users). The other happened at the arrival time from around 6:00 p.m. to around 12:00 a.m., marked by the red circle in Figure 3.10 (All users). A significant number of users have single activities that last for 9 to 14 hours in terms of stay duration, and arrival time spreads in between 6:00 p.m. to around 10:00 p.m. As a specific example, user #153 has another two different arrival times lasted for approximate 10 hours and 14 hours, marked by the red circles in Figure 3.10 (User #153).

Figure 3.10 The temporal patterns of stay duration and arrival time

Through analyzing the temporal patterns of single activities, some locations, such as home, office and gym, could be identified according to the pre-existing human behavioral regularities. For instance, in general, home and office/school are the two distinct locations where people spend longer hours (in average 8 - 10 hours) than that of the other activities, while the start time of these two types of places is very different. Normally, the time of arriving home is in the evening or at night since people usually go home directly after work/school or leisure activities, such as dining or fitness training in gyms. In contrast, the arrival time of office/school is expected to occur in the morning. Thus, the distribution of single activities in Figure 3.10 reflects the reality. There are two

time windows for arrivals, morning and evening to night, with stay duration of approximately 9 - 14 hours.

Since most activities are located in the north-west part of Beijing where several universities (e.g., Peking University, Tsinghua University, etc.) and research institutes (e.g., Microsoft Research Asia, Chinese Academy of Science, etc.) are located, which can infer that users in the GeoLife project worked in these universities and research institutes. For university students, faculty and staff, they travel in between different classrooms, dine in the cafeterias, and have recreations, such as soccer and basketball, at outdoor playground. Consequently, Figure 3.10 reflects high density of activities during time durations of 1 hour (for meals) and 2 hours (for class or recreations after classes). To summarize, some common characteristics are found by visually analyzing these activities such as the location of certain changed activities. The semantics of some places could be inferred through analyzing the temporal patterns of activities.

3.2.2 Activity clusters exploration

All users' single activities are clustered, from which top 5 sample clusters of two typical users (User #153 and User #128) are deleted based on the heavy-tail distribution of the number of single activities in each activity clusters for further analysis. All the single activities of these selected clusters are projected from WGS84 to UTM Zone 50N for visualizing them in 3D. Figure 3.11 shows the distribution of the selected activity clusters of User #153 and the activity distribution in each of these cluster. In the figure, the x-axis refers to the east direction in meter, whereas the y-axis refers to the north direction in meter. The specific (x, y) coordinates represent where a single activity occurred and the z-axis refers to the time of day when the single activity happened.

There are five activity clusters in Figure 3.11a. Each of them is zoomed in and showed in Figure 3.11b, 3.11c, 3.11d, 3.11e and 3.11f, where how the clustered single activities look like in time

and space can be examined. Three of these activity clusters (#1, #2 and #3) happened at the same location but in three different periods. It may infer that this user visited the same location three times per day. For example, some people leave home in the morning, return home at noon for lunch, and leave home again after lunch on weekdays; or she/he changed the activity (the time of visiting the place is changed). For instance, some bus drivers work in the night, followed by a day-shift later on. In this case, the location of bus pick-up is the same but the time is different. On the other hand, activity cluster #4 is likely similar to activity cluster #1 according to the method of detecting similar activity clusters proposed in this paper; thus, an activity change may have happened. Moreover, activity cluster #5 has high single activity density (Figure 11f), which means that user visited this place during certain time in a high probability.

Figure 3.11 Activity clusters of user #153 (graphs scaled based on the clusters): (a) five cluster samples; (b) - (f) single activities in cluster #1-5 shown in (a)

With regard to user #128, as shown in Figure 3.12a, no activity clusters seem to be at the same location but in different times. However, activity clusters #1 and #4 are found to be similar. These two activity clusters do not have high single activity density (Figure 3.12b and 3.12e), which means this type of activity did not happen very often. On the other hand, the activity clusters with the highest single activity density, i.e., activity clusters #3 and #5, do not have similar activity clusters; thus, it could infer that there is not any significant activity change happened. The whole mobility pattern should be more stable than that of User #153.

Figure 3.12 Activity clusters of user #128: (a) five cluster samples; (b) - (e) single activities in cluster #1-5 shown in (a)

Coincidentally, the activity cluster distributions of these two users match the single activity distributions shown in Figure 3.9b which has obvious activity change and in Figure 3.9c which always has the three identical activities in general.

3.2.3 Predictive results and evaluation

The activity clusters during January 2010 to June 2011, with and without implementing activity change detection (Algorithm 1), are used to train the predictive model. After implementing the predictive model, the activity clusters during July 2011 to September 2011 are used to evaluate the results. The transition probabilities between the activity clusters from the dataset during July 2011

to September 2011 is calculated, which are treated as the evaluation benchmark. The results of the predictive model without considering activity change detection are used to compare with the results that consider activity change detection.

Two activity clusters of A and B are selected and their predictive results and the real transition probabilities are listed in Table 3.2 for user #153. Comparing the predictive results between no activity detection and Algorithm 1, the transition probabilities from activity cluster A to B, D and F are set to 0 with conducting Algorithm 1, which well matches the real transition probability during July 2011 to September 2011. In other words, the activity cluster B, D and F are changed activity clusters. Although these activity clusters happened during the early period, even some activity clusters are visited in a very high frequency from activity cluster A (e.g., the highest transition probability from A to B), the changed activity cluster are successfully removed by the Algorithm 1 that reduced the negative impact on predictive results. However, a new activity cluster (H) is involved in the visited activity clusters after July 2011, which cannot be detected. In fact, this type of case happens only if the new activity cluster comes later after model training. On the other hand, a similar activity cluster, F (new), of activity cluster F is successfully detected. Therefore, the transition probability of F (new) is 0.19 in Algorithm 1 instead of 0.12. Based on the discussion of clustering results, it can infer that activity clusters F and F (new) correspond to activity clusters #1 and #4 in Figure 3.11 (a).

With respect to user #128, since the activity pattern of this user is stable overall, the results match the benchmark data very well even without performing change detection (Table 3.3). However, there still exist similar activity clusters with low single activity density (Figure 3.12c and 3.12f). In other words, for those users whose activities are not changed in general (refer to Figure 3.9c), they possibly still have activity change but only during a short period. Furthermore, the transition

from B to D in Table 3.3 only happened during July 2011 to September 2011, which is not part of the training dataset. Therefore, the transition is not detected by the predictive model based on the dataset from January 2010 to June 2011, which leads the predictive model to predict this transition in a wrong way.

The R^2 values of these two predictive results based on the transition probability from July 2011 to September 2011 are computed for both users #153 and #128. For the predictive results without conducting activity change detection, the R^2 is 0.295 for user #153, while the R^2 is 0.965 for user #128. For the predictive results with consideration of the Algorithm 1, the R^2 is 0.762 for user #153, while the R^2 is 0.971 for user #128. Apparently, the R^2 has an obvious improvement from 0.295 to 0.762, when Algorithm 1 is used to handle activity changes of human mobility. For users that have stable activity pattern, the predictions with and without implementing change detection both have a very good performance.

Table 3.2 Transition probability between activity clusters of user #153

Transitions	Jan 2010 to Jun 2011 (no activity change detection)	Jan 2010 to Jun 2011 (Algorithm 1)	Jul 2011 to Sept 2011
A -> B	0.38	0.00	0.00
A -> C	0.27	0.70	0.50
A -> D	0.19	0.00	0.00
A -> E	0.08	0.20	0.25
A -> F	0.04	0.00	0.00
A -> G	0.04	0.10	0.00
A -> H	0.00	0.00	0.25
B -> A	0.69	0.69	0.60
B -> F	0.12	0.00	0.00
B -> F(new)	0.07	0.19	0.20
B -> G	0.12	0.12	0.00
B -> H	0.00	0.00	0.20

Table 3.3 Transition probability between activity clusters of user #128

Transitions	Jan 2010 to Jun 2011 (no activity change detection)	Jan 2010 to Jun 2011 (Algorithm 1)	Jul 2011 to Sept 2011
A -> B	0.84	0.84	0.90
A -> C	0.16	0.16	0.10
B -> A	0.15	0.15	0.10
B -> C	0.75	0.75	0.70
B -> D	0.00	0.00	0.10
B -> E	0.05	0.00	0.00
B -> E(new)	0.05	0.10	0.10

3.3 Summary

In this chapter, an approach of human activity change detection was first proposed. Subsequently, a predictive model taking into account activity changes and built upon the concept of Markov chains was developed. The proposed predictive model considers the spatiotemporal impacts on Markov chain states, leading to a more precise prediction. The proposed predictive model was implemented for two users selected through visually analyzing the GeoLife dataset to evaluate the model performance. The results show that the R^2 value is improved from 0.295 to 0.762 for the user with obvious activity changes, and 0.965 to 0.971 for the users without obvious activity changes. Thus, the proposed methods, in terms of modelling activity changes, clustering activities and predicting human movement, present an effort towards effective ways of improving the accuracy in analyzing and predicting human movements.

However, to deal with such activity changes which may happen out of training samples, the key is to understand how a type of human mobility pattern forms. Therefore, it is essential to infer and uncover the motivations behind human mobility patterns, which is going to be discussed in the next chapter.

Chapter 4 Human Activity Motivation Modelling and Analysis

This chapter develops a spatiotemporal-semantic model based on Topic modelling to quantitatively analyze the relationship between the spatiotemporal patterns and the associated motivations in human activity patterns. The motivations behind spatiotemporal patterns of human activity patterns are inferred by the associated semantic patterns. Geo-tagged tweets posted in Toronto were used to analyze human activity patterns based on the spatiotemporal-semantic model, where a hypothesis—similar human mobility patterns may have different motivations—was proven.

4.1 Methodology

This section starts by defining two important terms that help understand the method described in the later sections. Topic models are then introduced, which provide basis for the discussion of spatiotemporal-semantic modelling for human activity patterns.

4.1.1 Preliminary definitions

Definition 4.1. A **stayed place**, p , is a place where there exist certain activities. A p can be represented as a set of triples $\{d_w, p_c, t_d\}$, where d_w , p_c and t_d refer to day of week, category of places and time of day, respectively. For example, a stayed place can be a restaurant where a person has a meal at 1:00pm on Monday. In this case, p can be indicated by $\{\text{Monday, Restaurant, 1:00pm}\}$.

Definition 4.2. **Semantics**, s , illustrates a type of non-spatial attribute about a person. The attributes may reflect a life style (e.g., jogging, vegetarianism, or commuting to work by car), an

event (e.g., attending a concert, watching a sport game, or participating a parade), a type of hobby (e.g., listening to music, doing sports, or reading books), a topic (e.g., U.S. president election, NBA final, or performance of a new smart phone), and so on, by means of topics in which people are interested in. Symbol s is defined as a distribution over words. For example, $s = \{\text{Lakers (0.3), nice (0.1), shot (0.2), Kobe (0.4)}\}$ indicates a topic about a nice L.A. Lakers' game where Kobe Bryant outperforms the game. Here it can infer that this person could have an attribute indicating that he is a fan of L.A. Lakers.

4.1.2 Probabilistic topic models

Topic modelling was originally proposed for understanding the topics in large text corpora within the domain of machine learning (Blei et al., 2003; Griffiths & Steyvers, 2002, 2003 & 2004; Hofmann, 1999 & 2001). The basic idea of Topic modelling is that documents are mixtures of topics (a probability distribution over topics), where a topic is a probability distribution over words. A topic model is a generative model specifying a simple probabilistic procedure. Documents can be generated by this probabilistic procedure. More specifically, a distribution over topics is initially chosen to make a new document. Then, for each word in that document, a topic is randomly chosen and a word can be drawn from that topic according to a distribution over words. This process can be inverted by using standard statistical techniques, which can infer the set of topics involved in generating a collection of documents and a set of words involved in generating a collection of the topics. Only words in documents can be observed (i.e., observed data). Based on certain probabilistic sampling rules, how words distributed in documents might be generated on the latent variables. These latent variables (probability distributions over topics and words) can then be found.

For the topic models described above, the generative process does not take into account the order of words appearing in documents. The number of times how a word appeared in documents is the only relevant information to the model, which is known as the *bag-of-words* assumption. This assumption is classic in statistical models of language processing (Steyvers & Griffiths, 2007). However, another topic model proposed by Griffiths et al. (2004) is sensitive to word order during the training procedure. It considers the syntactic as well as semantic factor to infer the topics of documents. In our study, the data of both spatiotemporal patterns and semantic patterns is trained without considering the sequence of “words”. On one hand, the stayed places are discrete data that are irrelevant to the order of each of the previous stayed places. On the other hand, the semantic information is represented by very short texts (e.g., 140 words in Tweets), not even a complete sentence, and sometimes just some key words, which can be considered as order insensitive.

Other than text mining, topic models have been applied to understanding activity patterns. For example, Huynh et al. (2008) utilized the topic model to discover daily routines based on wearable sensor data. Farrahi & Gatica (2011) applied the topic model on mobile phone data to explore routines. Hasan & Ukkusuri (2014) proposed a topic model-based method to classify urban activities using Twitter check-in data. However, the majority of these works emphasize only on modelling spatiotemporal patterns without further exploring the motivation of the patterns.

4.1.3 Spatiotemporal-semantic model for human activity patterns

A. Model description

A human activity pattern is represented by a set of two types of distributions. One is a distribution over sets of stayed places. The other is a distribution over the associated semantics (topics) from those sets of stayed places. A set of stayed places refers to a type of places, which is illustrated by the distribution over stayed places. $P(p/z)$ is used for the distribution over stayed places under a

set of stayed places (category) z . A distribution over sets of stayed place, $P(z)$, therefore, can describe the spatiotemporal pattern of a type of activity. A semantic pattern is defined by a distribution over topics. $P(s)$ is used to represent the associated semantic pattern, indicating the motivation behind the spatiotemporal pattern. Given a topic, $P(w/s)$, a distribution over words is used to describe what a topic is discussed about. As a result, a human activity pattern D , can be defined by:

$$D = \{P(z), P(s)\} \quad (4.1)$$

Only stayed place (p) represented by the triples and words in semantics (s) are visible in the model.

A graphical model (Figure 4.1), similar to many other studies, is drawn and used to further explain the process of activity pattern inferring. To simplify notations, it first assigns $\phi^{(n)}$ and $\phi'^{(n)}$ to $P(p/z = n)$ and $P(w/s = n)$ to represent the multinomial distribution over stayed places for a set of stayed places n and the multinomial distribution over words for topic n , respectively, and $\theta^{(m)}$ and $\theta'^{(m)}$ to $P(z)$ and $P(s)$ to represent the multinomial distribution over sets of stayed place for activity pattern m and the multinomial distribution over topics for activity pattern m , respectively. z and s in Figure 4.1 are the assignment of stayed place tokens to sets of stayed places and the assignment of word tokens to topics, respectively. The data $\phi^{(n)}$, $\phi'^{(n)}$, $\theta^{(m)}$, $\theta'^{(m)}$, z and s are latent variables that are required to be inferred. The box I and K indicate the repeated sampling of $\phi^{(n)}$ and $\phi'^{(n)}$ for each sets of stayed places z and topic s , respectively, until I sets of stayed places and K topics have been generated. The box J_m and L_m represent that, for each activity pattern, the repeated sampling of sets of stayed places / stayed places and topics / words, respectively, until J_m stayed

places and L_m words have been generated. The boxes M illustrate the sampling of a distribution over sets of stayed places and topics in each activity pattern m for total M patterns, respectively. The variables of α and β are hyper-parameters of Dirichlet prior, treated as constants. For an instruction of Dirichlet prior, please refer to the work of Blei et al. (2013). Consequently, in each activity pattern, the probability of the i^{th} stayed place and the i^{th} word in an activity pattern m are:

$$P(p_i|m) = \sum_{j=1}^I P(p_i|z_i = j)P(z_i = j|m) \quad (4.2)$$

$$P(w_i|m) = \sum_{j=1}^K P(w_i|s_i = j)P(s_i = j|m) \quad (4.3)$$

where z_i is a latent variable indicating the set of stayed places from which the i^{th} stayed place was drawn, s_i is a latent variable indicating the topic from which the i^{th} word was drawn. $P(p_i|z_i = j)$ and $P(w_i|s_i = j)$ indicate the probability of the i^{th} stayed place in set of stayed places j and i^{th} word for topic j , respectively; $P(z_i = j|m)$ and $P(s_i = j|m)$ refer to the probability of the j^{th} set of stayed places sampled for the i^{th} stayed place token and the probability of the j^{th} topic sampled for i^{th} word token for pattern m , respectively.

Figure 4.1 The topic model-based graphical models for describing spatiotemporal and semantic patterns. Shaded and unshaded circles represent observed and unobserved variables, respectively. Arrows refer to the dependences between variables. Boxes indicate repetitions of sampling steps.

Based on the above discussion, an example of human activity pattern structure is given in Figure 4.2. An activity pattern consists of a spatiotemporal pattern and a semantic pattern. With respect to the spatiotemporal pattern, from top to bottom, it is filled by three sets of stayed places with different probabilities (0.2, 0.8 and 0.0). For each set of stayed places, the appearing number of each stayed place varies. For example, in set of stayed places 1, p_3 appears three times while only shows once in set of stayed places 3. All the stayed places are from the stayed places database that is inputted from social media applications (e.g., Twitter, Facebook etc.). Moreover, the semantic pattern is represented by different topics with a variety of probability. In Figure 4.2, the three topics with probability of 0.8, 0.1 and 0.1 form the semantic pattern. The three topics contain different numbers of some specific words, which brings different meanings to those topics. All the words are generated from corresponding stayed places.

Figure 4.2 Structure of human activity pattern

B. Estimation of parameters

There has been a variety of techniques to estimate the parameters of topic models (Hofmann, 1999, Blei et al., 2003; Buntine, 2002; Griffiths & Steyvers, 2004; Hasan & Ukkusuri 2014). For our model, it avoids to directly estimate “stayed place – set of stayed places distributions” ϕ , word-topic distributions ϕ' , set of stayed places distributions θ and topic distributions θ' , since it may cause problems from estimating process (Steyvers & Griffiths, 2007). Instead, Gibbs sampling is used to directly estimate the two posterior (conditional) distributions over z (the assignment of stayed place token to sets of stayed places) and s (the assignment of word token to topics). Based on that, the spatiotemporal pattern (sets of stayed places distribution) and the semantic pattern (topics distribution) are easy to be estimated.

Gibbs sampling, a form of Markov chain Monte Carlo (MCMC), is used to do inference of topics because it is able to efficiently and easily extract topics from a large size data set (Steyvers & Griffiths, 2007). A simulation of a complex (often high-dimension) distribution can be conducted

by Gibbs sampling on lower-dimensional subsets of data, and all other data conditions each subset. More details about how this technique extracts topics can be found in Griffiths & Steyvers (2004).

In our model, the collection of spatiotemporal patterns is represented by, for each stayed place token i , a group of stayed place indices p_i and activity pattern indices m_i . The collection of semantic patterns is represented by, for each word token i , a group of word indices w_i and activity pattern indices m_i . For spatiotemporal patterns, each stayed place token is considered in turn, and the probability of assigning the current stayed place token to each set of stayed places is estimated, conditioned on the set of stayed places assignments to all other stayed place tokens. Based on this conditional distribution, a set of stayed places is sampled and forms the new set of stayed places assignment for this stayed place token. This conditional distribution (posterior distribution) can be written as $P(z_i = n | \mathbf{z}_{-i}, p_i, m_i, \cdot)$. According to Griffiths & Steyvers (2004), it can be calculated by:

$$P(z_i = n | \mathbf{z}_{-i}, p_i, m_i, \cdot) \propto \frac{C_{p_i n}^{PJ} + \beta}{\sum_{p=1}^P C_{pn}^{PJ} + P\beta} \frac{C_{m_i n}^{MJ} + \alpha}{\sum_{j=1}^J C_{m_i j}^{MJ} + J\alpha} \quad (4.4)$$

where $z_i = n$ indicates the set of stayed places assignment of token i to set of stayed places n , $\mathbf{z}_{-i}$ represents the set of stayed places assignments of all stayed place tokens except token i , and “ $\cdot$ ” refers to all other known or observed data. Moreover, C^{PJ} and C^{MJ} are matrices of counts with $P \times J$ and $M \times J$ dimensions, respectively. C_{pn}^{PJ} refers to the number of times, excluding the current instance i , stayed place p is assigned to set of stayed places n and C_{mn}^{MJ} refers to the number of times, excluding the current instance i , set of stayed places n is assigned to some stayed place token

in activity pattern m . As a result, the left part of the Equation (4.4) reflects the probability of stayed place p in set of stayed places n , while the part in the right hand side is the probability of set of stayed places n in activity pattern m . Similarly, for semantic patterns, each word token is considered in turn, and the probability of assigning the current word token to each topic is estimated, conditioned on the topic assignments to all other word tokens. Based on this conditional distribution, a topic is sampled and forms the new topic assignment for this word token. This conditional distribution (posterior distribution) can be written as $P(s_i = n | \mathbf{s}_{-i}, w_i, m_i, \cdot)$. According to Griffiths & Steyvers (2004), it can be calculated by:

$$P(s_i = n | \mathbf{s}_{-i}, w_i, m_i, \cdot) \propto \frac{C_{w_i n}^{WL} + \beta}{\sum_{w=1}^W C_{w n}^{WL} + W\beta} \frac{C_{m_i n}^{ML} + \alpha}{\sum_{l=1}^L C_{m_i l}^{ML} + L\alpha} \quad (4.5)$$

where $s_i = n$ indicates the topic assignment of token i to topic n , $\mathbf{s}_{-i}$ represents the topic assignments of all topic tokens except token i , and “ $\cdot$ ” refers to all other known or observed data. Moreover, C^{WL} and C^{ML} are matrices of counts with $W \times L$ and $M \times L$ dimensions, respectively. $C_{w n}^{WL}$ refers to the number of times, excluding the current instance i , word w is assigned to topic n and $C_{m n}^{ML}$ refers to the number of times, excluding the current instance i , topic n is assigned to some word token in activity pattern m . As a result, the left part of the Equation (4.5) reflects the probability of word w in topic n , while the part in the right hand side is the probability of topic n in activity pattern m .

Based on the conditional distributions, stayed place – set of stayed places distribution ($\hat{\phi}$) and spatiotemporal pattern ($\hat{\theta}$) can be estimated by:

$$\hat{\phi}_i^{(j)} = \frac{C_{ij}^{PJ} + \beta}{\sum_{k=1}^P C_{kj}^{PJ} + P\beta} \quad (4.6)$$

$$\hat{\theta}_j^{(m)} = \frac{C_{mj}^{MJ} + \alpha}{\sum_{k=1}^J C_{mk}^{MJ} + J\alpha} \quad (4.7)$$

And, word – topic distribution ($\hat{\phi}'$) and semantic pattern ($\hat{\theta}'$) can be estimated by:

$$\hat{\phi}'_i^{(j)} = \frac{C_{ij}^{WL} + \beta}{\sum_{k=1}^W C_{kj}^{WL} + W\beta} \quad (4.8)$$

$$\hat{\theta}'_j^{(m)} = \frac{C_{mj}^{ML} + \alpha}{\sum_{k=1}^L C_{mk}^{ML} + L\alpha} \quad (4.9)$$

C. Estimation of number of topics

The number of topics need to be estimated in advance for model training. The perplexity, used by convention in machine learning, is a standard metric to measure the generalization performance (Blei et al., 2003), which was used to estimate the number of topics. The perplexity presents an exponentially decreasing trend in the likelihood of the training data, which is used to measure the performance of a probability model. The lower the perplexity value, the better generalization

performance of the model. In our case, for a training dataset of M users, the perplexity in spatiotemporal dataset (D_{st}) is defined as:

$$Perplexity(D_{st}) = \exp\left\{-\frac{\sum_{m=1}^M \log P(\mathbf{p}_m|M)}{\sum_{m=1}^M J_m}\right\} \quad (4.10)$$

where J_m is the number of stayed places in each user, $\mathbf{p}_m$ refers to stayed places of user $m \in D_{st}$. $P(\mathbf{p}_m|M)$ can be computed by Equation (4.2). The perplexity in semantics dataset (D_{sem}) is defined as:

$$Perplexity(D_{sem}) = \exp\left\{-\frac{\sum_{m=1}^M \log P(\mathbf{w}_m|M)}{\sum_{m=1}^M L_m}\right\} \quad (4.11)$$

where L_m is the number of words of tweets of each user, $\mathbf{w}_m$ refers to words of user $m \in D_{sem}$. $P(\mathbf{w}_m|M)$ can be computed by Equation (4.3).

4.2 Results and discussion

The methodology was implemented based on the tweets posted within Toronto, Canada. Both spatiotemporal pattern and semantic pattern of each user were determined. Subsequently, the similarity was calculated for both patterns between users to further explore these patterns.

4.2.1 Data

A crawler program was developed to collect tweets posted in Toronto using Twitter Streaming APIs (<https://dev.twitter.com/streaming/overview>). Over five million (5,081,215) tweets posted by 77,386 users from March 10, 2014 to November 14, 2014 were obtained. The users who averagely posted at least one tweet per day were selected to make sure all users in the case study are daily active users, no further filtering was done. As a result, a total of 1,748,623 tweets from 4,174 users were obtained. Each collected tweet contains UserID, Time (yyy-mm-dd, hh:mm:ss), Geo-coordinates (longitude and latitude coordinates), Text, and so on. Time and Geo-coordinates were used to infer spatiotemporal patterns, while Text was used to understand semantic patterns.

To set up the stayed place according to Definition 4.1, it is needed to convert geo-coordinates to categories of places, land-use (Figure 4.3) is used to categorize the places where tweets posted. The land-use data was created in 2011 with the resolution of 30 meters. In this case, the error of the geo-coordinates was removed and the geospatial information of tweets were represented by categories of land-use where some specific activities occurred. Moreover, days of a week are represented by “1” to “7” indicating Monday to Sunday, respectively, and time of day was converted to minutes. Eventually, a stayed place, which was used to train model, was formed as “ d_{wpcld} ” such as “1residential1480”, “5commercial1139” and “6openarea748”. “1residential1480” refers to an activity occurs at a residential place on Monday around midnight; “5commercial1139” refers to an activity occurs at a commercial area on Friday around 6pm; and “6openarea748” refers to an activity occurs at an open area on Saturday around noon. With respect of text of tweets, only letters (case insensitive) and numbers were used to form tokens in the model training. As such, each user was treated as a document where words are replaced by “ d_{wpcld} ”, and letters and numbers.

Figure 4.3 Toronto land-use category (Obtained from City of Toronto)

4.2.2 Pattern inferring

The model is developed based on a machine learning toolkit, MALLET (McCallum, 2002), using Java in Eclipse. For estimating the number of topics, both spatiotemporal dataset and text dataset are split into two parts randomly, 80% of the dataset was used for training while the rest of it was used to compute the values of perplexity according to Equation (4.10) & (4.11). Figure 4.4 illustrates the distribution of perplexity values over the number of sets of stayed places and topics trained by the spatiotemporal pattern model and the semantic pattern model. For spatiotemporal pattern model, 40 sets of stayed places were used to represent the spatiotemporal pattern for each user since the perplexity on other numbers after 40 shows insignificant change (Figure 4.4 left), whereas there is no significant decreasing of perplexity for semantic pattern model. And the perplexity is going up once the number of topics is increased by 350 (Figure 4.4 right). In this case, it is difficult to tell which number is good for inferring the topics of semantic patterns. 120 and 350 are used to do inferring because the model is relatively stable between 120 and 350 topics. The inferred topics are manually checked, and found that there are more topics consisting of similar

words in the pattern with 350 topics than the pattern with 120 topics. As a result, the semantic pattern with 120 topics makes more sense.

In order to understand the rationale why the perplexity values of semantic pattern model fit this form of distribution, the text of tweets is further checked. Comparing with the spatiotemporal dataset that each unit data is written in the specific format (“*dwpctd*”), the words of text in tweets are much more diverse. Furthermore, individuals use a variety of special letter combinations to express some topics with the same meaning. For example, any of “haha”, “lol” and “lool” words may be used to express “laugh out loud”. In other words, it forces the model to generate new topics to actually express the same meaning.

Figure 4.4 Perplexity results on spatiotemporal pattern model (left) and semantic pattern model (right)

A total of 4,174 users’ activity patterns represented by both spatiotemporal patterns and semantic patterns in Toronto were inferred. For each user, the spatiotemporal pattern is represented by the distribution over 40 sets of stayed places, each set of stayed places is represented by the distribution over stayed places that is able to explain the spatiotemporal pattern of this set of stayed places.

The semantic pattern is represented by the distribution over 120 topics; each topic is described by the distribution over “words” in the tweets.

Table 4.1 shows the top 10 distributions over stayed places in 3 of 40 sets of stayed places. These distributions reveal some patterns. For example, the pattern of set of stayed places #4 reveals that the activities happened in residential areas mainly in the morning at around 10:00am on Wednesday, Thursday and Saturday and at night around 10:00pm every day. Set of stayed places #32 reveals another type of pattern that should be referred to certain activities happened in government or institutional around 8:00pm. For the pattern of set of stayed places #38, it can tell that leisure activities in bar or nightclub are more likely involved since parks or recreational are the place category where some activities happened around midnight.

Table 4.1 The distribution over stayed places for sets of stayed places

Set of stayed places #4		Set of stayed places #32		Set of stayed places #38	
Stayed place	P(p z)	Stayed place	P(p z)	Stayed place	P(p z)
5residential1380	0.006	4government&institutional481	0.016	6parks&recreational79	0.003
3residential659	0.005	7government&institutional481	0.014	5parks&recreational1411	0.002
4residential1319	0.004	6government&institutional481	0.013	3parks&recreational74	0.002
6residential660	0.004	4government&institutional491	0.012	7parks&recreational0	0.002
2residential1320	0.004	4government&institutional482	0.010	1parks&recreational116	0.002
7residential1320	0.004	3government&institutional481	0.009	1parks&recreational59	0.002
4residential659	0.004	3government&institutional491	0.009	4parks&recreational115	0.002
2residential0	0.004	5government&institutional481	0.009	4parks&recreational94	0.002
1residential1379	0.004	7government&institutional482	0.008	3parks&recreational981	0.002
4residential1380	0.004	7government&institutional490	0.008	5parks&recreational38	0.002

A total of six topics with top 10 distributions over words are also selected to illustrate some semantic patterns (Table 4.2). It is obvious that the probabilities on words in these six topics except for Topic #99 and Topic #64 are rather low overall, which is probably because more words are involved leading to lower probability allocated to each word. However, for the topic-itself, the rank of words should be paid more attention regardless how low the probability on each word is. Topic #9 talks about bars, parks, restaurants and coffee house in Toronto; Topic #33 talks about

fitness, Topic #36 talks about eating a pizza in a pizza store, Topic #64 talks about listening a music played by a street artist, Topic #87 talks about watching a music video in Toronto, and Topic #112 refers to a topic about food in different areas (Mississauga, Brampton and Oakville). Moreover, some other topics involving in only adjectives (e.g., happy, sad, beautiful, sexy, cute, love, et al.) do exist, which describe a type of mood or attitude. The combination of different topics may denote different semantic patterns.

Table 4.2 The distribution over words for topics

Topic #9		Topic #33		Topic #36	
Word	P(w/s)	Word	P(w/s)	Word	P(w/s)
toronto	0.147	gym	0.025	pizza	0.022
Bar	0.013	fitness	0.019	chicken	0.014
centre	0.011	workout	0.016	pizzapizza	0.013
ontario	0.011	work	0.009	Pop	0.009
Park	0.01	fit	0.009	Store	0.007
St	0.009	goodlife	0.007	dundas	0.007
restaurant	0.008	health	0.007	toronto	0.006
canada	0.007	yoga	0.007	italian	0.005
house	0.007	jeffthehotbet	0.007	Cheap	0.005
coffee	0.007	app	0.006	hungary	0.005
Topic #64		Topic #87		Topic #112	
Word	P(w/s)	Word	P(w/s)	Word	P(w/s)
music	0.122	watch	0.021	food	0.026
artist	0.084	toronto	0.016	mississauga	0.015
noise	0.015	pvr	0.011	excellent	0.014
street	0.013	video	0.011	village	0.011
drumming	0.009	event	0.009	pizza	0.007
hot	0.007	music	0.008	grill	0.007
yonge	0.007	canada	0.005	brampton	0.006
toronto	0.005	yyzhilighztz	0.005	oakville	0.006
st	0.004	2chicks1car	0.004	cafe	0.005
square	0.004	follow	0.004	district	0.005

Based on these two distributions, the activity pattern of each user is illustrated. Two users, user #1 and user #2, are randomly selected, to discuss their activity patterns using Table 4.3. Regarding user #1, he or she has two popular types of stayed places, set of stayed places #8 and #33, where

some activities happened with the motivation mainly involving topic #36, #64 and #107. For user #2, although he or she has the similar motivation of his or her activities (high probability of topic #36 and #64 are involved in semantic patterns as well), the spatiotemporal pattern is different. To further analyze those activity patterns, the similarities between users are computed, which is going to be discussed in the next section.

Table 4.3 An example of activity patterns of two users

User #1			
Spatiotemporal pattern		Semantic pattern	
Set of stayed places	P(z)	Topic	P(s)
8	0.642	36	0.443
33	0.18	64	0.216
35	0.048	107	0.129
23	0.038	5	0.092
17	0.034	11	0.046
39	0.012	20	0.02
24	0.011	117	0.015
1	0.009	7	0.011
5	0.007	60	0.01
Others	Others	Others	Others
User #2			
Spatiotemporal pattern		Semantic pattern	
Set of stayed places	P(z)	Topic	P(s)
20	0.206	36	0.355
0	0.201	64	0.291
6	0.142	25	0.179
5	0.049	117	0.063
10	0.048	5	0.05
22	0.048	20	0.021
30	0.043	11	0.014
2	0.038	54	0.013
3	0.035	94	0.006
Others	Others	Others	Others

4.2.3 Analysis of activity pattern similarity

In our case study, the semantic pattern associated with the spatiotemporal pattern from tweets plays a critical role in the motivation of the spatiotemporal pattern. To find whether there are any

individuals that have similar spatiotemporal patterns but different motivations behind, the similarity in terms of spatiotemporal patterns and semantic patterns between pairwise users are computed. A standard function to measure the difference or divergence between two topic distributions p and q in two corresponding documents is Kullback Leibler (KL) divergence (Lin, 1991):

$$KL(p, q) = \sum_{i=1}^T p_i \log_2 \frac{p_i}{q_i} \quad (4.12)$$

where T refers to the number of topics. $KL(p, q)$ is equal to zero when the two documents are the same. Thus, two sets of stayed places from two users are used to replace p and q to compute the similarity in terms of spatiotemporal pattern. The topics from the text of tweets are used to replace p and q to compute the similarity in terms of semantic pattern. Since the KL divergence is asymmetric, for convenient, a symmetric measure is computed using:

$$KL' (p, q) = \frac{1}{2} [KL(p, q) + KL(q, p)] \quad (4.13)$$

KL' is computed between every two users, if the KL' of both spatiotemporal pattern and semantic pattern closer to zero together, the activity pattern of two users is more similar. Visually, if the values of KL' between one user and any other user are plotted in Figure 4.5, the pairwise users situated in more white area have higher similarities in terms of activity pattern.

Figure 4.5 Activity pattern similarity

Eight users are randomly selected to plot their KL' values in Figure 4.6. Overall, they all have a lot of other users with low KL' values in terms of semantic pattern, while a few of users having low KL' values in terms of spatiotemporal pattern are found. More specifically, points rarely appear in the areas nearing zero in user #3, #4 and #5. In contrast, user #8 and #10 have a bunch of users of whom the KL' values more close to 0. In other words, user #8 and #10 have users with higher similarity in terms of activity pattern than user #3, #4 and #5. Moreover, there exist some users that have similar spatiotemporal patterns but significantly different semantic patterns (the points close to the vertical axis but far away from the horizontal axis). For example, those users who are different in terms of semantic patterns but similar in terms of spatiotemporal patterns from user #8 can be found.

Figure 4.6 Distribution pattern of KL' values

Furthermore, statistics of KL' values of every two users are performed for all 4,174 users in Toronto. The KL' values were divided into seven groups: $[0, 0.5)$, $[0.5, 1)$, $[1, 1.5)$, $[1.5, 2)$, $[2, 2.5)$, $[2.5, 3)$ and $[3, +\infty)$, the number of user pairs were counted for each group. The user pairs in the first group have the highest similarity while have the lowest similarity in the last group. Figure 4.7 (left) shows the number of user pairs in each group in terms of spatiotemporal pattern and semantic pattern, respectively. The part with most population (37.37%) (with KL' values in $[0, 0.5)$) has the most similar semantic patterns, which means they have very similar topics, while the smallest part of users (6.24%) have the highest similarity in terms of spatiotemporal pattern. And, they are linked together to check the similarity in terms of both spatiotemporal and semantic patterns, which is plotted in Figure 4.7 (right). Only a very small number of users, 2.72%, have similar activity patterns with the KL' value in $[0, 0.5)$, while a majority of users, 87.14%, show significantly different activity patterns (similar spatiotemporal patterns and different semantic patterns, similar semantic patterns and different spatiotemporal patterns, or different in both).

Apparently, the population of users with very similar activity patterns, i.e., KL' value in $[0, 0.5)$, is decreased by 56.41% after adding semantics dimensions in spatiotemporal patterns.

Figure 4.7 The number of user pairs in different KL' value groups. The number of user pairs in different KL' value groups in terms of spatiotemporal pattern and semantic pattern, respectively (left) and the number of user pairs in different KL' value groups in terms of both spatiotemporal pattern and semantic pattern (right)

4.2.4 Discussion

As the hypothesis proposed at the beginning, for individuals who have similar mobility patterns, their activity patterns may be different since their hidden motivation varies. In this case, it is biased to find two similar individuals in terms of activity pattern by only looking into their spatiotemporal patterns. Indeed, it can be visually justified by the points situated close to the vertical axis but far away from the horizontal axis in Figure 4.6 (high similarity in terms of spatiotemporal pattern with very low similarity in terms of semantic pattern). Meanwhile, the population of the pairwise users having very high similar activity patterns, i.e., KL' value in terms of both spatiotemporal and semantic patterns in $[0, 0.5)$, is 56.41% less than the population of the pairwise users only having very high similar spatiotemporal patterns, i.e., KL' value in terms of only spatiotemporal pattern

in $[0, 0.5)$. It can further quantitatively illustrate that the similarity in terms of only spatiotemporal dimensions cannot align to the similarity of activity patterns.

More specifically, one user (User #11) who has a very similar spatiotemporal patterns (0.05 of KL' value in terms of spatiotemporal pattern) but very different semantic patterns (2.57 of KL' value in terms of semantic pattern) with user #10 is randomly selected to further explain. The spatiotemporal patterns with the sets of stayed places that have high visiting probability are listed in Table 4.4, and the corresponding motivation (semantic pattern) with the topics representing a type of activity is given in Table 4.5. In Table 4.4, it can see that both user #10 and #11 have set of stayed places #22 and #0. The stayed place distribution in both sets of stayed places are examined. In set of stayed places #22, the activities occurred at the places of resource and industrial land-use around 9:00 am on Wednesday, Thursday, Friday or Saturday. For set of stayed places #0, it involves the places of resource and industrial area where activities happened between 1:00 pm to 6:00 pm on Wednesday, Thursday, Friday or Saturday. However, set of stayed places #29 and #12 are a group residential places where activities mainly happened on Saturday and Sunday at around 9:00 am and midnight. Obviously, these two users have the similar spatiotemporal patterns where some activities occurred.

Table 4.4 Similar spatiotemporal patterns between two users

Spatiotemporal pattern			
User #10		User #11	
Set of stayed places	P(z)	Set of stayed places	P(z)
22	0.382	0	0.351
0	0.263	11	0.289
29	0.173	22	0.260
Others	Others	Others	Others

Table 4.5 Two different semantic patterns associated with the spatiotemporal pattern in Table 4.4

Semantic pattern			
User #10		User #11	
Topic	P(s)	Topic	P(s)
83	0.003	65	0.001
9	0.012	54	0.001
31	0.014	11	0.002
94	0.007	41	0.021
Others	Others	Others	Others

However, the associated semantic patterns are significantly different with respect to the topic distributions. The distribution over words for these topics is looked into. User #10 mainly involves the topics about public subway stations and bus stops in Scarborough and Etobicoke of Toronto (Topic #83), bars and restaurant (Topic #9), a good diner (Topic #31) and Toronto mayor election (Topic #94). In contrast, user #11 has the topics about beer, wine and cellar (Topic #65), Nintendo video games in Mississauga of Toronto (Topic #54), a goal in a Word Cup game (Topic #11), hiring a security to patrol and guard (Topic #41).

Based on both patterns, it can infer that user #10 visits some industry and resources places because of taking public transit since there are resource and industrial area surrounding subway stations and bus stops in both Scarborough and Etobicoke in Toronto, while user #11 may be a manager of an alcohol warehouse since he or she has the activity to hire securities to patrol and guard a specific cellar (Topic #65&41) happened in resource and industrial area. Apparently, the motivation of visiting industry and resources places for these two users is very different. Moreover, although both users have activities occurred at residential area, what they do could be significantly different as well. For user #10, he or she may watch morning news about Toronto mayor election, while user #11 properly plays video games or watches Word Cup game at night. Other activities may happen at other different types of places.

With the understanding of the motivation behind a type of mobility pattern, similar mobility patterns can be further distinguished and precisely predicting related events becomes feasible. For example, in our previous study (Huang et al., 2015), human activity changes could be modelled more reasonably, which enable human mobility to be predicted precisely even without updating predictive model before training new samples. Moreover, some events distributed in different locations across the city could be sensed that may contribute to emergency response.

The land-use dataset used for the case study involves seven categories. If the land-use dataset with more classifications are used, the spatiotemporal patterns would be more diverse, which may lead to less similar spatiotemporal patterns and activity patterns but with higher spatial similarities. Spatiotemporal clustering can be used to form spatiotemporal patterns, if the exacted locations need to be known for some applications.

4.3 Summary

This chapter presents a spatiotemporal-semantic model through using Topic modelling to analyze the relationship between the spatiotemporal patterns and the associated motivations in human activity patterns. A case study was conducted based on geo-tagged Twitter data collected within Toronto, Canada. Through computing the similarity between pairwise users with respect to their spatiotemporal pattern, semantic pattern and spatiotemporal-semantic pattern, the findings demonstrate that only a very small number of users (2.72%) have similar activity patterns with the KL' value in $[0, 0.5)$, while a majority of users (87.14%) show different activity patterns (i.e., similar spatiotemporal patterns and different semantic patterns, similar semantic patterns and different spatiotemporal patterns, or different in both). The population of the users having very similar activity patterns are decreased by 56.41% after considering semantic patterns in the

corresponding spatiotemporal patterns, which can prove the hypothesis—similar mobility patterns may have different motivations. By analyzing the activity patterns of the two selected users as an example, it indicates that similar spatiotemporal patterns can be further distinguished by incorporating the associated motivations. In other words, the motivation plays a key role in modelling and analyzing human activity patterns. The following chapter further models and analyzes human activity patterns from space, time and semantics.

Chapter 5 Human Activity Pattern Modelling and Analysis

In this chapter, human activity patterns are further explored and modelled based on not only spatiotemporal dimensions but also the associated semantic patterns representing the motivation—what people do, think and discuss in a location at a specific time. Spatiotemporal clustering was used to analyze the spatiotemporal pattern of human activities, and a machine learning method was used to infer the associated semantic pattern. Then the similarity between human activities was computed. Based on the similarity, a network consisting of human activities was ultimately constructed to represent human activity patterns. A whole year geo-tagged tweets posted within Toronto, Canada were used to analyze the human activity patterns.

5.1 Methodology

In this experiment, geo-tagged tweets were used to explore human activities. The textual content within the tweets should somehow reflect different topics occurred in different places and times, and thus offer a way to sense what individuals do, think and discuss. Each tweet is considered to represent an activity. Spatiotemporal clustering is used to analyze where and when a type of activity occurs, while the associated topics being discussed in each cluster are inferred based on Topic modelling. Through computing similarities among clusters in terms of time and semantics (topics), the human activity patterns can then be detected.

5.1.1 Data

The whole year of geo-tagged tweets posted in Toronto, Canada (April 2014 to April 2015) were downloaded through Twitter Streaming API, including 4,350,005 tweets from 125,730 users. The

Twitter data was first preprocessed to remove 1) the tweets that were sent from identical locations considered as robot accounts, and 2) the tweets sent by users whose annual average tweeting is less than the annual average of all users. The annual average of all users, 34.5 tweets in this case, was determined by a heavy-tail distribution shown in Figure 5.1a, since the mean value can represent the characteristics of the majority populations (Jiang et al., 2013).

After data filtering, 3,684,980 tweets from 18,122 users were selected for the study. Figure 5.1b illustrates the distribution of the selected tweets, displayed as 10 m by 10 m raster density map, where higher density of tweets is shown with brighter color. These tweets were anonymized so that only UserID (a unique number), Time (yyy-mm-dd, hh:mm:ss), Geo-coordinates (longitude and latitude coordinates) and Text were included in the study. Furthermore, the textual content of these tweets were filtered to keep only nouns, verbs and adjectives for semantic pattern analysis.

(a)

(b)

Figure 5.1 (a) Distribution of tweets' volume posted by each user; and (b) mapped tweets (higher density of tweets are displayed with brighter color)

5.1.2 Spatiotemporal clustering

Spatiotemporal clusters are the foundation for studying the human activity patterns with consideration of semantics that is going to be discussed in Section 5.1.4. Clustering in space and time was analyzed separately using DBSCAN method, because it does not require predetermination of the number of clusters and it is widely used in situations where large datasets need to be processed (Zhou et al., 2000). Different from other studies where both spatial and temporal clusters are analyzed together using some spatiotemporal clustering methods (e.g., Birant & Kut, 2007), clustering is performed in space and time sequentially, i.e., activities are grouped based on their location and then grouped the activities clustered around the same location based on time. Such clustering approach allows the final spatiotemporal clusters identically geo-referenced with the spatial clusters, which is essential for the subsequent steps in this study. Figure 5.2 illustrates the concept of the abovementioned clustering analysis.

Figure 5.2 Spatiotemporal activity clusters

The key of the clustering process is to define both distance and density parameters for spatial and temporal clustering. The error of GPS receivers equipped in smart phones most likely contribute to the shift of activities generated from an identical place, which can be used as a spatial distance threshold. Since the geo-location error of geotagged tweets has been literately reported within 10 meters (Soliman et al., 2015), the distance threshold is assigned as 10 meters. The density for the

spatial clustering can be estimated by the average number of activities in each 10 m by 10 m square since the number of activities in each square follows a type of heavy-tail distribution, which means the activity density of the most of the squares are around the average number of activities. With regards to the temporal clustering, a 30-minute threshold is used as the average time interval of revisiting a place is less than 30 minutes usually (Zhao et al., 2008) indicating that similar activities may be generated every 30 minutes or less. Since temporal clustering is conducted after spatial clustering, the density can be set as 1 to cover every possible similar activity in terms of occurring time.

5.1.3 Semantic pattern modelling

A semantic pattern is described by the distribution over a topic. The probabilistic distribution of topics in each spatiotemporal cluster is inferred to understand the associated activity patterns. Topic modeling is used to infer the distribution, the details about how to create a topic model and how to estimate the necessary parameters for the topic model can be found in Chapter 4.1.2.

5.1.4 Similarity among spatiotemporal activity clusters

The similarity between human activities is quantitatively analyzed based on not only the spatiotemporal dimensions but also semantic patterns extracted in Section 5.1.3. In order to do so, it defines that a group of spatiotemporal activity clusters are similar when they have close time period and similar semantic pattern. Figure 5.4 illustrates an example that cluster A and B are similar since they have the same time period (the height of cluster) and a similar semantic pattern (the color of cluster). A Boolean matrix, $A(s)_{n \times n}$ is used to quantitatively measure the similarity between every two spatiotemporal activity clusters:

$$A(S)_{n \times n} = A(TS)_{n \times n} \wedge A(SS)_{n \times n} = \begin{bmatrix} 0 & \cdots & ts_{0n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \wedge \begin{bmatrix} 0 & \cdots & ss_{0n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & s_{0n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \quad (5.6)$$

where $A(S)_{n \times n}$ refers to a n by n upper triangular matrix representing the similarity between every two spatiotemporal activity clusters among n spatiotemporal clusters, where the entries of main diagonal are all zero. The entry s_{kj} is assigned as 1, if cluster k and j are similar in terms of semantic and time dimensions, otherwise s_{kj} is assigned to 0, where $k, j \in [0, n]$. The same for $A(TS)_{n \times n}$, time similarity matrix, and $A(SS)_{n \times n}$, semantic pattern similar matrix. Then, $s_{kj} = ts_{kj} \wedge ss_{kj}$.

Figure 5.3 Similar activity clusters. The clusters having similar semantic pattern are highlighted in the same color. Therefore, cluster A and B are similar, while others are different

To determine “similarity” in terms of semantic and time dimensions, Kullback Leibler (KL) divergence, a standard function to measure the difference or divergence between two topic distributions p and q in two corresponding documents (Lin, 1991), is used to compute the similarity:

$$KL(p, q) = \sum_{i=1}^T p_i \log_2 \frac{p_i}{q_i} \quad (5.7)$$

where T refers to the number of topics. $KL(p, q)$ is equal to zero when the two documents are identical. In our case, the topics distribution from two activity clusters are assigned to p and q , respectively. Since the KL divergence is asymmetric, for convenience, a symmetric measure is computed using:

$$KL' (p, q) = \frac{1}{2} [KL(p, q) + KL(q, p)] \quad (5.8)$$

KL' between every two users is computed to further group similar activity clusters in terms of semantic pattern. It assumes that the two spatiotemporal clusters are not significant different if the value of KL' of them is less than 0.5, then the associated entry of $A(SS)$ in Equation (5.6) is assigned to 1, otherwise 0.

A general way to determine if activity cluster happened in the same duration can be achieved by defining certain time periods, such as 7:00 am to 7:00 pm and 7:00pm to 7:00 am (e.g., Mathew et al., 2012). However, such an approach leads to a low resolution in time dimension which may cause certain negative impacts in subsequent activity pattern analysis. A data-driven method based on Welch's t-test (Welch, 1947) then is applied to compute the similarity in terms of time period for clusters in this study. By using it, fixed time periods do not need to be defined in advance, where a diversity of time periods can be dynamically generated based on the characteristics of dataset for computing the similarity.

The statistic t of Welch's t-test is defined as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \quad (5.9)$$

where $\overline{X}_1, s_1^2$ and N_1 and $\overline{X}_2, s_2^2$ and N_2 are sample mean, sample variance and sample size of two datasets, respectively. Based on t , P -value is computed to finally determine if the two datasets are similar or not. Welch's t-test is applied to compare activity distribution in terms of time for every two clusters, therefore similar clusters have a characteristic that the majority of activities in each cluster of them are all distributed in the same time period. The entry of $A(TS)$ in Equation (5.6) is assigned to 1 if the P -value is greater than 0.95, otherwise 0.

5.2 Results and discussion

5.2.1 Spatiotemporal patterns

A total of 1,569 spatiotemporal clusters was generated from 423 different activity hotspots (spatial clusters). The distribution of the number of activity in each spatiotemporal cluster is a type of heavy-tail distribution (Figure 5.5), so only those spatiotemporal clusters with the number of activities greater than the mean are considered as meaningful samples for further analysis. Eventually, only 448 spatiotemporal clusters were selected for further study.

Figure 5.4 Distribution over tweet number in spatiotemporal clusters

For the selected spatiotemporal clusters, the number in each hotspot ranges from 1 to 25 (Figure 5.6a. Big circles may contain different types of activity, while the circles with the same size in different locations may be caused by the same type of activities.), while most of hotspots have less than 5 spatiotemporal clusters (Figure 5.6b). The mean of the number of spatiotemporal clusters in each hotspot is close to the corresponding median of the spatiotemporal clusters (Figure 5.6c), which is 3.7 and 4.0, respectively. In other words, each hotspot is averagely visited during 3.7 different time periods, which is probably driven by different types of activities. Moreover, the spatiotemporal clusters that have the same number of time periods (the same size of circles in Figure 5.6a) are geo-distributed diversely, which is likely caused by the same type of activities at different locations.

Figure 5.5 Spatiotemporal cluster distribution. (a) Mapped spatiotemporal clusters. (b) The histogram of the allocation of spatiotemporal clusters in terms of space. (c) The density of spatiotemporal clusters in each spatial cluster

Figure 5.7 shows the temporal pattern of activities in each hotspot (spatial cluster) as a vertical line, the lighter the color is, the more activities occurred during that time period. Various temporal patterns are drawn: some hotspots are visited in a similar frequency all the time (the vertical lines having similar color change patterns), while a number of places have significantly different visiting

frequencies (the vertical lines having different color change patterns). Moreover, a few locations which are frequently revisited in a specific time period are identified (e.g., cluster around 60 having only a bright-yellow block). Such hotspots, having significant color change, contains several spatiotemporal clusters, which can be considered as a place involved with different types of activities. Those vertical lines with similar color change patterns likely indicate similar types of activities. Another observation from the graph is that the activities between 4pm and 8pm is not that active overall.

Figure 5.6 Activity temporal pattern in different spatial clusters. Spatial clusters are ranked based on the associated number of activities

5.2.2 Semantic patterns

The topic model was developed based on a machine learning toolkit, MALLET (McCallum, 2002), written by Java. To estimate the number of topics, the text information associated with spatiotemporal clusters were randomly split into two parts, 80% of them were used for training

while the rest of the data were used to compute the value of perplexity according to Equation (5.5). A proper topic number can be used if the value of perplexity of other topic numbers decreases flatly afterward. As a result, a total of 100 topics was used to train the topic model although the value of perplexity looks more stable after 160, since the content of 100 topics has less redundancy than 160 topics (Figure 5.8). The value of α is assigned as 0.01 and the value of β can be computed by $50/T = 0.5$ (T refers to the number of topics) since they are reasonable choices for topic models (Stein & Griffiths, 2007).

Figure 5.7 Perplexity distribution over topic number

The semantic pattern in each spatiotemporal cluster is quantitatively represented by distribution over the inferred 100 topics. Table 5.1 gives an example of the distribution for a spatiotemporal cluster (only top 10 topics are listed) randomly selected. Obviously, there are eight topics involved in this semantic pattern, topic #56 is the most popular topic among them. Table 5.2 illustrates top 10 word distributions in topic #56, which discusses about Toronto Raptors (an NBA team) and its stadium (Air Canada Centre). Based on these key words, it can infer that it could be an activity

related to NBA games about Toronto Raptors. The time occurred is around 7:00 pm and the activity is found around a residential building, which can give us a sense that it could be an activity about Raptors' fans watching the basketball TV game rather than staying in the stadium or discussing the team's performance after the game. Similarly, the same semantic pattern associated with the spatiotemporal cluster around Air Canada Centre could be an activity about watching a Toronto Raptor's game in the stadium. Although the two activities happened in different locations, they can be seen as one type of activities, watching NBA games, if taking semantics into account.

Table 5.1 Semantic pattern of a selected spatiotemporal cluster as an example

Topic ID	Distribution (P(s))
56	0.803
87	0.119
37	0.034
66	0.025
13	0.008
33	0.008
91	0.003
75	0.001
0	0.000
1	0.000

Table 5.2 Top 10 word distributions in topic 56

Word	Distribution (P(w s))
Canada	0.068
Air	0.066
Ventre	0.061
Raptors	0.042
Toronto	0.032
Game	0.021
Air Canada centre	0.020
We the north	0.017
Leafs	0.013
Ontario	0.012

5.2.3 Activity patterns

The activity pattern is finally drawn by connected spatiotemporal clusters (Figure 5.9), a type of network. The connections stand for the similarity in terms of both semantics and time, which is computed by Equation (5.6). In other words, the connected spatiotemporal clusters make up a type of activities regardless how far they are away each other geographically.

Figure 5.8 Activity patterns depicted as a network

In Figure 5.9, there exist many locations having significant spatial difference but are similar in terms of semantics and time pattern (connected circles at different locations). This pattern illustrates that people could go to different places for the same purpose or some similar events could happen at different locations. More importantly, the locations that are far away from each other but connected directly (by red lines) could actually affect each other (The color of lines indicates the distance between two connected activities. The distance ranges from 20 m to 48,459 m.). For example, if the mobility pattern of a specific crowd in one area (A) changed due to a certain reason caused by participating in activities, the mobility pattern of individuals in another

area (B), whose associated activities connect to the activities of the crowd in area A , most likely changes in the future. If this mobility pattern change may cause a certain negative effect for the communities, such as increasing the traffic, some measures can be taken in advance in the area B so as to avoid relevant problems. On the other hand, if a certain group of people, such as the people often takes outdoor activities, in location A can be affected by certain factors, like air pollution, which subsequently moves to location B , it can predict that location B 's individuals who have connected activities with the people in location A are most likely being affected. Consequently, these connected activities can serve as a component for predicting some similar events happening in different locations.

The connectivity of each cluster ranges from 1 to 59, representing the size of circles in Figure 5.9. Low connectivity refers to weak impact in terms of semantic and temporal dimensions, which means that the activity has low probability to be affected by or to affect other activities, while high connectivity can be seen as a type of unstable status since it has higher chance to be impacted and also may cause more activities to be affected. On the other hand, such activities with low connectivity can be seen as a type of special or irregular activity, while frequently connected activities are more routine. These locations associated with the activities having low connectivity should be given more attention since they have higher probability to generate events in cities.

Moreover, a large number of spatiotemporal clusters with very similar semantic patterns exist (Figure 5.10. If the number of KL' is between 0 and 0.5, the two involved spatiotemporal clusters are considered to have similar semantic pattern. In this regard, a value of 1 is assigned to the corresponding entry of matrix $A(SS)$ in Equation (5.6).), where 55% of them contributes to KL' , of which the value is less than 0.5. This can quantitatively provide an evidence that people may participate in similar activities in different locations due to the similar functionalities involved in

different urban areas. In other words, some different mobility patterns actually are driven by similar motivations.

Figure 5.9 Distribution of number of spatiotemporal clusters among KL'

5.3 Discussion

With the extracted activity pattern, a corresponding activity network was created. By using this network, human mobility can be reasonable predicted. More possible places to where individuals are moving could be provided by means of analyzing the connected activities. More importantly, activity changes can be predicted in advance through finding those similarly connected activities in case where an activity change is detected. As such, these activities can be removed from prediction model to decrease the negative impact of the predictive results. Furthermore, through analyzing the structure of associated activity network of a type of mobility pattern from different individuals, the mobility pattern could be further distinguished.

The relationship between connectivity of activity and the number of individuals associated with each different connectivity of activity are further plotted (Figure 5.11a). Several spikes at low connectivity and high connectivity can be observed. The higher the spikes reach, the larger the population associates to a connectivity. Based on that, a Gaussian mixture model can be used to locate the statistically low / high connectivity so as to analyze their corresponding populations. Finally, values with 12 and 21 were used to represent low connectivity and high connectivity, respectively (Figure 5.11b). With this setting, there exist about 10% of individuals having activities with low connectivity, while about 61% people have activities with high connectivity. In other words, most of people in cities repeat the same activities daily, and they live based on a specific activity pattern. In contrast, only a very small group of people have carry out a variety of activities, where this group of people has a living style tend to be “independent” from the mass. Our findings also echo the argument raised by Gonzalez et al. (2008) that individuals follow simple reproducible mobility patterns. Also, our work further proves that human movement can be quantitatively predicted. On the other hand, it can be used as an indicator to measure the “personality” of a city. An active city should have higher population with lower connectivity, and vice versa.

Figure 5.10 Connectivity distribution over population. (a) Relationship between connectivity and the number of individuals. (b) Population distribution over connectivity by Gaussian Mixture Model (GMM)

The importance of understanding the interaction of urban environments and urban citizens for improving urban life has been raised by Albeverio et al. (2008). Although the advances in pervasive computing have provided rich and solid data foundation, proper models capable of harnessing such data need to be further explored to accumulate our understandings. The network created in this study, depicting activity patterns, can be used as a base to develop more complex models. This study used only a type of data; however, richer activities can be involved by following the same approach and coupling more diverse datasets (i.e., biometric data). Through analyzing the network structure of the activities involving individuals, communities and cities, the people, communities and cities can be further categorized. Through analyzing the relationship between the networks associated with the individuals, communities and cities, how they influence (interact) each other could be inferred. In other words, the interaction between micro objects (e.g., people) and macro objects (e.g., urban environments) could be simulated based on the network, which may better illustrate complex urban systems.

5.4 Summary

This chapter further analyzes human activity patterns from spatiotemporal dimensions and the related motivation inferred by semantic patterns as addressed in Chapter 4. The DBSACN clustering method was implemented to analyze the spatiotemporal patterns, and Topic modelling was used to infer the associated semantic patterns. The similarity between spatiotemporal clusters of human activities was computed, which was used to build up a network to ultimately represent human activity patterns. One-year geo-tagged tweets in Toronto was processed and used to analyze intra-urban human activity patterns, the result reports that 61% of individuals live based on a specific activity pattern, while only 10% of individuals have a living style tend to be “independent” from the mass. In addition, through analyzing the network structure of connected activities derived by individuals, communities and cities, as well as the relationship within those networks, how the networks evolve from a micro scale to a macro scale can be simulated, which can better understand how complex urban systems work.

Chapter 6 Conclusions and Future Work

6.1 Conclusions

How people live, work and play in cities play a key role in a variety of urban applications, as such it is essential to understand human activity patterns. Predicting human movements and modelling human activity patterns are the two main perspectives that researchers place an emphasis on. One of the challenges limiting the predictive accuracy of human movements is the negative impact of activity changes. Therefore, in this thesis, an approach was first developed to detect activity changes, of which a predictive model was developed based on Markov chain. The predictive model was tested using massive GPS trajectories of human movements provided by the Microsoft GeoLife project. The results show that the R^2 is improved from 0.295 to 0.762 for the user who has significant activity changes and 0.965 to 0.971 for the user without obvious activity changes. The proposed methods involved in the predictive model achieve the first research objective as indicated in Section 1.3.

On the other hand, most of the existing researches regarding modelling human activity patterns mainly focus on spatiotemporal dimensions, the motivation laid behind is often neglected. A spatiotemporal-semantic model was developed based on Topic modelling to depict how people live, work and play in cities, with emphasize on modelling the associated motivation behind spatiotemporal patterns of human activities. Semantic patterns were used to represent what people do and discuss and thus infer the motivation behind spatiotemporal patterns. Through applying the proposed spatiotemporal-semantic model on geo-tagged tweets posted within Toronto, Canada, the relationship between the spatiotemporal patterns and the associated motivations in human activity patterns was quantitatively analyzed. The experiments found that the population of those

users having very similar spatiotemporal patterns are decreased by 56.41% after considering the associated semantic patterns. The findings provide scientific support to a hypothesis, where similar human mobility patterns may have different motivations, and address the importance of determining the motivations. In addition, it can further quantitatively illustrate that determining the similarity only using spatiotemporal dimensions cannot align to the similarity of activity patterns. By using the developed model, the similar spatiotemporal patterns can be further distinguished by incorporating the associated motivations. The developed spatiotemporal-semantic model and the results have successfully met the second research objective in Section 1.3.

Through building on top of the spatiotemporal-semantic model, a network model was created to present human activity patterns in cities. The network consists of similar activities in terms of time and semantics. A whole-year geo-tagged tweets posted within Toronto, Canada was used to analyze the human activity patterns, the results report that there exist 61% of individuals living in a regular pattern, while only 10% of individuals have a living style tend to be “independent” from the mass. In addition, through analyzing the network structure of connected activities derived by individuals, communities and cities, as well as the relationship within those networks, how the networks evolve from a micro scale to a macro scale can be simulated, which can better understand how complex urban systems work. The aforementioned studies on modelling and understanding human activity patterns achieve the third research objective in Section 1.3.

6.2 Future work

In the near future, this thesis’s work can be extended to further explore how urban citizens interact with urban environments by incorporating more types of relevant datasets. For instance, with the emergence of personal wearable devices, continuous measurements of biometric data such as

blood glucose levels, blood pressure and heart beat rate can be collected so that the information can be manipulated to understand the health condition and how they influence the movement, or vice versa. Once the relationship is built between the movement and biometric reaction, a further enhanced version, named “spatiotemporal-semantic-biometric” model can be constructed for better uncovering the human mobility and activity. As well, a more universal approach of model validation and accuracy assessment should be designed and discussed by the research community so as to provide a benchmark for the advent of different predictive models.

On the other hand, new ways are expected to be explored to predict the evolution of the entire urban system by means of analyzing individuals' activity changes. More specifically, regardless of individuals or city clusters, any “similar objects” may affect each other through various ways. Among a group of similar objects, if the state of one of the activities changes, most of them are likely to be involved in their subsequent changes. For example, if a state of a city changes, cities with similar activity patterns would be affected in a way where their state change may occur later on. The reason causing this state change may be due to a state change of a community in the city, while the state change of a community may be triggered by a state change of a group of similar activities. If we further look into the state change of these similar activities, it may be caused by a specific activity change as a result of human movement. Consequently, a state change of a large scale object (e.g., a city) could be inferred at a very early stage according to the state change of any micro scale object (e.g., an activity) by means of detecting change being occurred.

The aforementioned process is very similar to the phenomenon of “Butterfly Effect”—any small changes may have critical impacts on a dynamic system, and such argument may also occur in urban systems. If a human activity network can be formed based on certain similarities which are computed from various human mobility and activity datasets, this type of “Butterfly Effect”

application can be further explored by analyzing the evolution of network structure from micro scale to macro scale before and after certain events happened. Some related urban applications may be benefited, for example, certain diseases may be caused by attending a certain type of activities. If such diseases are detected from one community, the communities involving similar activity patterns should be paid with more attention for preventing the outbreak of the detected diseases. In addition, the affected communities would be inferred and predicted by only detecting and analyzing the activities involved in the diseases. It would be very useful for preventing such epidemic when its pathogeny has not been fully elucidated.

References

- Alexander, L., Jiang, S., Murga, M., & González, M. C. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, 240-250, 2015.
- Ashbrook, D., & Starner, T. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5), 275-286, 2003.
- Albeverio, S., Andrey, D., Giordano, P., & Vancheri, A. The dynamics of complex urban systems. *Physica, Heidelberg*, 2008.
- Asahara, A., Maruyama, K., Sato, A., & Seto, K. Pedestrian-movement prediction based on mixed Markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 25-33, New York, NY, USA, November 1-4, 2011.
- Austin, D., Cross, R. M., Hayes, T., & Kaye, J. Regularity and predictability of human mobility in personal space. *PloS one*, 9(2), e90256, 2014.
- Becker, H., Naaman, M., & Gravano, L. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*, pp. 438-441, Barcelona, Spain, July 17-21, 2011.
- Belik, V., Geisel, T., & Brockmann, D. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1(1), 011001, 2011.
- Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. Sentiment in new york city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*, 2013.

Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022, 2003.

Birenboim, A., & Shoval, N. Mobility Research in the Age of the Smartphone. *Annals of the American Association of Geographers*, 106(2), 283-291, 2015.

Birant, D., & Kut, A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1), 208-221, 2007.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pp. 427-434, Istanbul, Turkey, November 12-16, 2014.

Brockmann, D., Hufnagel, L., & Geisel, T. The scaling laws of human travel. *Nature*, 439(7075), 462-465, 2006.

Buntine, W. Variational extensions to EM and multinomial PCA. In *Machine Learning: ECML 2002*, pp. 23-34, Springer Berlin Heidelberg, 2002.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A. L. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22), 224015, 2008.

Carrothers, G. A. An historical review of the gravity and potential concepts of human interaction. *Analytical human geography*, American Elsevier Publishing Co., New York, USA, 1969.

Chen, J., Shaw, S. L., Yu, H., Lu, F., Chai, Y., & Jia, Q. Exploratory data analysis of activity diary data: a space-time GIS approach. *Journal of Transport Geography*, 19(3), 394-404, 2011.

Dietterich, T. G. Machine learning for sequential data: A review. In Proceedings of Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 15-30. Berlin, Germany, August 6, 2002.

Eagle, N., Pentland, A. S., & Lazer, D. Inferring friendship network structure by using mobile phone data. In Proceedings of the National Academy of Sciences, 106(36), 15274-15278, 2009.

Erlander, S., & Stewart, N. F. The gravity model in transportation analysis: theory and extensions (Vol. 3). Vsp, 1990.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226-231, Portland, Oregon, August 2-4, 1996.

Etter, V., Kafsi, M., Kazemi, E., Grossglauser, M., & Thiran, P. Where to go from here? Mobility prediction from instantaneous information. *Pervasive and Mobile Computing*, 9(6), 784-797, 2013.

Fast, S. M., González, M. C., Wilson, J. M., & Markuzon, N. Modelling the propagation of social response during a disease outbreak. *Journal of the Royal Society Interface*, 12(104), 20141105, 2015.

Farrahi, K., & Gatica-Perez, D. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1), 3, 2011.

Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications*, 41(13), 5733-5750, 2014.

Frias-Martinez, E., Williamson, G., & Frias-Martinez, V. An agent-based model of epidemic spread using human mobility and social network information. In Proceedings of in Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 57-64, Boston, MA, USA, October 9-11, 2011.

Gambs, S., Killijian, M. O., & del Prado Cortez, M. N. Next place prediction using mobility Markov chains. In Proceedings of the ACM First Workshop on Measurement, Privacy, and Mobility, pp. 1-6, Bern, Switzerland, April 10-13, 2012.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782, 2008.

Griffiths, T., & Steyvers, M. A probabilistic approach to semantic representation. In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp. 381-386, Fairfax, Virginia, USA, August 7-10, 2002.

Griffiths, T., & Steyvers, M. Prediction and semantic association. In *Advances in Neural Information Processing Systems* 16, 11-18, 2003.

Griffiths, T. L., & Steyvers, M. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1), 5228-5235, 2004.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), 260-271, 2014.

Hanson, S. Perspectives on the geographic stability and mobility of people in cities. *Proceedings of the National Academy of Sciences*, 102(43), 15301-15306, 2005.

- Hasan, S., & Ukkusuri, S. V. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44, 363-381, 2014.
- Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57, Berkeley, California, USA August 15-19, 1999.
- Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2), 177-196, 2001.
- Huang, W., Li, S., Liu, X., & Ban, Y. Predicting human mobility with activity changes. *International Journal of Geographical Information Science*, 29(9), 1569-1587, 2015.
- Huang, W. & Li, S. Understanding human activity pattern based on space-time-semantics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016 (Accepted).
- Huynh, T., Fritz, M., & Schiele, B. Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pp. 10-19, Seoul, Korea, September 21-24, 2008.
- Jeung, H., Liu, Q., Shen, H. T., & Zhou, X. A hybrid predictive model for moving objects. *IEEE 24th International Conference on Data Engineering*, pp. 70-79, Cancun, Mexico, April 7-12, 2008.
- Jiang, B., Liu, X., & Jia, T. Scaling of geographic space as a universal rule for map generalization. *Annals of the Association of American Geographers*, 103(4), 844-855, 2013.
- Joh, C. H., Arentze, T., Hofman, F., & Timmermans, H. Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research Part B: Methodological*, 36(5), 385-403, 2002.

Kang, C., Ma, X., Tong, D., & Liu, Y. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4), 1702-1717, 2012.

Kwan, M. P. Gender, the home-work link, and space-time patterns of non-employment activities. *Economic Geography*, 75(4), 370-394, 1999.

Liao, L. Location-based activity recognition (Doctoral dissertation, University of Washington), 2006.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... & Cheng, T. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119-133, 2016.

Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145-151, 1991.

Lu, E. C., & Tseng, V. S. Mining cluster-based mobile sequential patterns in location-based service environments. In *Proceedings of Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pp. 273-278, Taipei, Taiwan, May 18-20, 2009.

Mathew, W., Raposo, R., & Martins, B. Predicting future locations with hidden Markov models. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 911-918, Pittsburgh, Pennsylvania, USA, September 5-8, 2012.

Markov, A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain, 1971.

Merler, S., & Ajelli, M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1681), 557-565, 2010.

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." <http://mallet.cs.umass.edu>. 2002.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. WhereNext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 637-646, Paris, France, June 28-July 1, 2009.

Morzy, M. Mining frequent trajectories of moving objects for location prediction. In *Machine Learning and Data Mining in Pattern Recognition*, pp. 667-680, Springer Berlin Heidelberg, 2007.

Ossama, O., & Mokhtar, H. M. Similarity Search in Moving Object Trajectories. In *Proceedings of the 15th International Conference on Management of Data*, pp. 1-6, Mysore, India, December 9-12, 2009.

Park, M. H., Hong, J. H., & Cho, S. B. Location-based recommendation system using Bayesian user's preference model in mobile devices. In *Ubiquitous Intelligence and Computing*, pp. 1130-1139, Springer Berlin Heidelberg, 2007.

Pak, A., & Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), pp. 19-21, Valletta, Malta, May 17-23, 2010.

Pan, B., Zheng, Y., Wilkie, D., & Shahabi, C. Crowd sensing of traffic anomalies based on human mobility and social media. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 344-353, Orlando, Florida, USA, November 5-8, 2013.

Poletto, C., Tizzoni, M., & Colizza, V. Human mobility and time spent at destination: impact on spatial epidemic spreading. *Journal of Theoretical Biology*, 338, 41-58, 2013.

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5), 727-748, 2006.

Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286, 1989.

Riedel, D. E., Venkatesh, S., & Liu, W. Recognising online spatial activities using a bioinformatics inspired sequence alignment approach. *Pattern Recognition*, 41(11), 3481-3492, 2008.

Ravenstein, E. G. The laws of migration. *Journal of the Statistical Society of London*, 48(2), 167-235, 1885.

Sakaki, T., Okazaki, M., & Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, pp. 851-860, Raleigh, North Carolina, USA, April 26-30, 2010.

Sen, A., & Smith, T. Gravity models of spatial interaction behavior. Springer Science & Business Media, 2012.

- Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. Unravelling daily human mobility motifs. *Journal of the Royal Society Interface*, 10(84), 20130246, 2013.
- Sutton, C., & McCallum, A. An introduction to conditional random fields. *arXiv preprint arXiv:1011.4088*, 2010.
- Song, C., Qu, Z., Blumm, N., & Barabási, A. L. Limits of predictability in human mobility. *Science*, 327(5968), 1018-1021, 2010.
- Song, L., Kotz, D., Jain, R., & He, X. Evaluating next-cell predictors with extensive Wi-Fi mobility data. *Mobile Computing, IEEE Transactions on*, 5(12), 1633-1649, 2006.
- Soliman, A., Yin, J., Soltani, K., Padmanabhan, A., & Wang, S. Where Chicagoans tweet the most: Semantic analysis of preferential return locations of Twitter users. *arXiv preprint arXiv:1512.06880*, 2015.
- Simini, F., González, M. C., Maritan, A., & Barabási, A. L. A universal model for mobility and migration patterns. *Nature*, 484(7392), 96-100, 2012.
- Steyvers, M., & Griffiths, T. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7), 424-440, 2007.
- Stouffer, S. A. Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 5(6), 845-867, 1940.
- Takeuchi, Y., & Sugimoto, M. CityVoyager: an outdoor recommendation system based on user location history. In *Ubiquitous Intelligence and Computing*, pp. 625-636, Springer Berlin Heidelberg, 2006.
- Tversky, A. Features of similarity. *Psychologica Review* 84, 327–352, 1997.

Vail, D. L., Veloso, M. M., & Lafferty, J. D. Conditional random fields for activity recognition. In Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 235, Honolulu, HI, USA, May 14-18, 2007.

Wang, D., & Song, C. Impact of human mobility on social networks. *Journal of Communications and Networks*, 17(2), 100-109, 2015.

Welch, B. L. The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2), 28-35, 1947.

Work, D., Bayen, A., & Jacobson, Q. Automotive cyber physical systems in the context of human mobility. In Proceedings of National Workshop on High-Confidence Automotive Cyber-Physical Systems, Troy, MI. Troy, MI, April 3-4, 2008.

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one*, 9(5), e97010, 2014.

Xia, Y., Wang, G. Y., Zhang, X., Kim, G. B., & Bae, H. Y. Spatiotemporal similarity measure for network constrained trajectory data. *International Journal of Computational Intelligence Systems*, 4(5), 1070-1079, 2011.

Ying, J. J. C., Lee, W. C., Weng, T. C., & Tseng, V. S. Semantic trajectory mining for location prediction. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 34-43, Chicago, Illinois, USA, November 1-4 2011.

Yuan, Y., & Raubal, M. Measuring similarity of mobile phone user trajectories—a Spatiotemporal Edit Distance method. *International Journal of Geographical Information Science*, 28(3), 496-520, 2014.

Zhao, M., Mason, L., & Wang, W. Empirical study on human mobility for mobile wireless networks. In Proceedings of Military Communications Conference, pp. 1-7, San Diego, California, USA, November 17-19, 2008.

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y. Understanding mobility based on GPS data. In Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 312-321, Seoul, Korea, September 21-24, 2008.

Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th International Conference on World Wide Web, pp. 791-800, Madrid, Spain, April 20-24, 2009.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W. Y. Recommending friends and locations based on individual location history. ACM Transactions on the Web (TWEB), 5(1), 5, 2011.

Zhou, A., Zhou, S., Cao, J., Fan, Y., & Hu, Y. Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of Computer Science and Technology, 15(6), 509-526, 2000.