Fusing Social Media and Traditional Traffic Data for Advanced Traveler Information and Travel Behavior Analysis

by

Zhenhua Zhang
January 15, 2017

A dissertation submitted to the Faculty of the Graduate School of the University at Buffalo, the State University of New York in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Civil, Structural, and Environmental Engineering
ACKNOWLEDGEMENTS

First, I would like to offer my greatest gratitude to my supervisor, Dr. Qing He, who has supported me throughout my research studies showing their patience and knowledge. His continuous encouragement and guidance on my research works in the past three and a half years help me with useful suggestions to make research plans. Without his help, this dissertation would not have been possible.

At the same time, I would give my thankfulness to the committee members: Dr. Adel W. Sadek and Dr. Qian Wang, who accept the invitation to serve on my committee and give valuable devices on my dissertation. Besides, thank them a lot for teaching me a lot of useful knowledge about transportation modelling and demand forcasting in their class.

In addition, I thank for the research environment provided by the University at Buffalo in which the students can devote themselves to the research. I can concentrate on my studies all the time without worries about affairs, which are irrelevant to research. Thank very much Ming Ni, who provides the social media data for my research. Thank my friends Lei Lin, Yu Cui and Li Tang who helped me a lot in my previous studies.

Finally, thank my family for giving me the material and spiritual help. Their unconditioned care always pushes me forward. Hope they would be happy for my Ph.D. graduation.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................................................. II

LIST OF FIGURES ............................................................................................ IX

LIST OF TABLES .................................................................................................. XV

ABSTRACT ........................................................................................................... 1

CHAPTER 1 INTRODUCTION ............................................................................. 5

1.1 Research background and motivations ....................................................... 5

1.1.1 Research background ........................................................................... 5

1.1.2 Motivations ............................................................................................ 8

1.2 Data description .......................................................................................... 9

1.2.1 Loop detector data .............................................................................. 9

1.2.2 Social media data ................................................................................. 11

1.2.3 Traffic accident data ............................................................................ 13

1.2.4 Land use data and Google Place Type data ........................................... 14

1.2.5 Trip data of connected vehicle .............................................................. 14

1.4 Contributions ............................................................................................ 15

1.5 Dissertation organization .......................................................................... 18

CHAPTER 2 LITERATURE REVIEW .................................................................. 21

2.1 Regional traffic flow pattern identification ............................................... 21

2.1.1 Gaps in large-scale traffic pattern identification .................................... 21
2.1.2 Pattern identification and anomaly detection ............................................. 22
2.1.3 Dimensionality reductions of traffic data .................................................... 25
2.2 Social media definitions and features ........................................................... 26
2.3 Travel behavior and travel motivation identification ......................................... 30
  2.3.1 Mobility features of travel behavior based on Twitter ................................. 33
2.2.3 Travel motivation identification ................................................................ 36
2.2.4 Research gaps and opportunities ............................................................... 38
2.3 Traffic accident and social event detection ..................................................... 40
  2.3.1 Traffic accident detection based on social media ........................................ 40
  2.3.2 Traffic surge and social event detection based on social media .................. 45
2.4 Trip purpose inference .................................................................................. 49

CHAPTER 3 SPATIAL-TEMPORAL TRAFFIC FLOW PATTERN IDENTIFICATION
AND ANOMALY DETECTION ................................................................................. 55

  3.1 Data preprocessing ....................................................................................... 55
  3.2 Method ......................................................................................................... 58
    3.2.1 Dictionary-based compression .................................................................. 58
    3.2.2 Minimum Description Length principle ................................................. 61
    3.2.3 Anomaly degree .................................................................................... 63
  3.3 Numerical examples .................................................................................... 67
    3.3.1 Spatial regional traffic flow pattern ....................................................... 67
3.3.2 Temporal regional traffic flow pattern ............................................. 72

3.4 A case study .......................................................................................... 76

3.5 Comparisons......................................................................................... 81

3.5.1 Comparisons between different discretization settings .................... 81

3.5.2 Comparisons with another anomaly detection method..................... 84

3.6 Conclusions and discussions ................................................................ 86

CHAPTER 4 EXPLORING TRAVEL BEHAVIOR: ABNORMAL MOVEMENTS USING HIGH-RESOLUTION TWEET TRAJECTORY .................................................. 89

4.1 Empirical findings of tweet trajectory ................................................. 89

4.1.1 Location features ............................................................................ 89

4.1.2 Movement features ........................................................................ 92

4.1.3 Clustering features ........................................................................ 96

4.2 Geo-Mobility method .......................................................................... 98

4.2.1 Geo clustering................................................................................... 98

4.2.2 Abnormal movement detection ..................................................... 103

4.3 Travel motivation decoding ............................................................... 106

4.4 Conclusions and discussions .............................................................. 109

CHAPTER 5 THE HUMAN MOBILITY PATTERN STUDY BASED ON SOCIAL MEDIA .............................................................................................................. 112

5.1 Human mobility study ........................................................................... 112
5.1.1 Twitter displacement and human mobility patterns .................................. 112
5.1.2 Long-distance Twitter displacement ................................................................ 118
5.2 Travel motivation inference ................................................................................ 121
  5.2.1 Topic generation ......................................................................................... 121
  5.2.2 Topic tokens and their interpretations .......................................................... 125
5.3 Applications and discussions ............................................................................. 129
  5.3.1 Individual mobility pattern and travel motivation identification .............. 130
  5.3.2 Social event influence identification ............................................................ 132
5.4 Conclusions and discussions ............................................................................. 135

CHAPTER 6 TRAFFIC ACCIDENT DETECTION WITH BOTH SOCIAL MEDIA AND
TRAFFIC DATA .................................................................................................... 138

  6.1 Data preprocessing ......................................................................................... 138
    6.1.1 Tweet data preprocessing ..................................................................... 138
    6.1.2 Token filtering and stemming ............................................................... 139
  6.2 Classification by SVMs .................................................................................. 141
    6.2.1 The entire process ................................................................................ 141
    6.2.2 Classification with individual word ....................................................... 142
    6.2.3 Classification with paired words ........................................................... 145
  6.3 Comparing with the classification by sLDA .................................................. 150
    6.3.1 sLDA Process ...................................................................................... 150

vi
6.3.2 Comparisons of classification results .......................................................... 151

6.4 Improvements of classification by traffic-related information ....................... 152

6.4.1 Flow-occupancy pattern modeling ................................................................. 153

6.4.2 Abnormal pattern identification and traffic-related features ....................... 157

6.4.3 Classification results with traffic-related information ................................. 159

6.5 Comparisons with the ground truth ................................................................. 160

6.5.1 Coverage of accident-related tweets ............................................................. 160

6.5.2 Features of time and space differences between tweets and accidents ....... 161

6.5.3 Features of unrecorded tweet accidents ....................................................... 163

6.6 Conclusions and discussions ........................................................................... 164

CHAPTER 7 AN EXPLORATORY STUDY ON THE CORRELATION BETWEEN TWITTER CONCENTRATION AND TRAFFIC SURGE ......................................................... 167

7.1 Twitter concentration study method ................................................................. 167

7.1.1 Twitter concentration extraction and filtering ............................................. 167

7.1.2 Twitter concentration classification and labeling ....................................... 170

7.2 Traffic surge detection method ...................................................................... 173

7.2.1 Traffic clustering algorithm ........................................................................ 173

7.2.2 Traffic surge definition in detector level .................................................... 177

7.3 Correlations between Twitter concentrations and traffic operations .......... 178

7.4 Conclusions and discussions ........................................................................... 183
CHAPTER 8 AN EMPIRICAL STUDY OF TRIP PURPOSE INFERENCE WITH CONNECTED VEHICLES TRAJECTORIES, LAND USE DATA, AND SOCIAL MEDIA DATA .................................................................................................................................................................................. 185

8.1 Data preprocessing and empirical findings ................................................................. 185
8.1.1 Trip information from tweets .................................................................................. 185
8.1.2 Trip information from Connected-Vehicle ............................................................... 187
8.1.3 Problem statement and study flow ........................................................................... 189
8.2 Trip purpose inference ............................................................................................. 190
8.2.1 Tweet location extraction ...................................................................................... 190
8.2.2 Trip purpose inference with tweet locations ........................................................... 193
8.2.3 Comparisons with land use data ............................................................................. 196
8.3 Conclusions and discussions ..................................................................................... 198

CHAPTER 9 DISSERTATION CONCLUSION AND FUTURE SCOPE ............................ 201

9.1 Conclusions ............................................................................................................. 201
9.2 Future scope ........................................................................................................... 205

REFERENCE .................................................................................................................... 210
LIST OF FIGURES

Figure 1-1 (a) Locations of the intersections (white dots), and the boundaries of sub-regions, and (b) Layouts of the detectors in signalized intersections ........................................ 11

Figure 1-2 Study area for tweets in (a) NOVA and (b) City of Ann Arbor..................... 12

Figure 1-3 Paper organization.......................................................................................... 19

Figure 2-1 546 Twitter locations of a person at 17:00 p.m. in different weekdays. The locations are visually clustered where the size of Cluster 1 is 481, Cluster 2 is 19 and Cluster 3 is 21. ............................................................................................................................ 36

Figure 2-2 Recurrent traffic fluctuations of hourly traffic occupancy in one sample detector........................................................................................................................................ 46

Figure 2-3 The comparison between land use and tweets in inferring the trip purposes. ........................................................................................................................................... 51

Figure 3-1 Three geographic scales to identify both the spatial and temporal traffic flow pattern within a road network .......................................................................................... 56

Figure 3-2 spatial regional traffic flow pattern identification during PM peak (5:00-6:00 p.m.) on (a) Feb 20, 2014 (Weekday) and (b) Mar 22, 2014 (Weekend). The anomaly degrees of all detectors (x-axis of the bar plots) in two most abnormal intersections and their locations are shown in the bottom of the plot......................................................... 69

Figure 3-3 Anomaly degrees of spatial traffic flow patterns of two intersections on PM peak (5:00-6:00 p.m.) in different days. The first intersection (a) is Richmond Hwy. and Sherwood Hall Ln.; the second intersection (b) is Lee Hwy. and ramp of W Ox Rd. ..... 71

Figure 3-4 Spatial anomaly degrees in three different time periods for different sub-regions (each link represents a sub-region). ......................................................... 72
Figure 3-5 Temporal regional traffic flow pattern identification during PM peak (5:00-6:00 p.m.) on (a) Feb 20, 2014 (Weekday) and (b) Mar 22, 2014 (Weekend). The anomaly degrees of all detectors (x-axis of the bar plots) in two most abnormal intersections and their locations are shown in the bottom of the plot.

Figure 3-6 Anomaly degrees of temporal traffic flow patterns of two intersections on PM peak (5:00-6:00 p.m.) in different days. The first intersection (a) is Seminary Rd. and South George Mason Rd.; the second intersection (b) is Rixlew Ln. and Wellington.

Figure 3-7 Temporal anomaly degrees in three different time periods for different sub-regions (each link represents a sub-region).

Figure 3-8 Spatial and temporal regional traffic flow pattern of all detectors on Feb 20, 2014 (Weekday) (a) 07:00-08:00 a.m., (b) 12:00-13:00 p.m. and (c) 17:00-18:00 p.m. and Mar 22, 2014 (Weekend) (d) 07:00-08:00 a.m., (e) 12:00-13:00 p.m. and (f) 17:00-18:00 p.m.

Figure 3-9 Map of the sub-region; Ratios of spatial anomaly degrees between after and before Sliver Line extension in the same area (x-axis and y-axis are longitude and latitude, respectively) on (b) weekdays and (c) weekends; Ratio of temporal anomaly degrees between after and before Sliver Line extension on (d) weekdays and (e) weekends.

Figure 3-10 Examples of anomaly degrees in one detector during 17:00~17:15 p.m. on different weekdays. Obvious abnormal dots are circled. The boundary settings are defined as follows, (a) the same setting as introduced in Section 3; (b) different levels when discretizing traffic flow to [0, 500), [500, 100), [1000, 2000]; (c) different levels when discretizing traffic occupancy to [0, 20%), [20%, 40%), [40%, 60%), [60%, 80%), [80%, 100%].
Figure 3-11 Examples of anomaly degrees in two detectors during 17:00~17:15 p.m. on different weekdays with different boundary settings. The results of the first detector are shown in (a)-(c) and those of the second are shown in (d)-(f). (a) and (d) follow the boundary setting [0, 1/17, 1/9, 1/7, 1/2, 1] as introduced in Section 3; (b) and (e) merge some occupancy categories into [0, 1/17, 1/7, 1]; (c) and (f) further divide the occupancy categories into [0, 1/28, 1/17, 1/12, 1/9, 1/7, 1/2, 1].

Figure 3-12 Temporal anomaly results by the proposed compression-based method, (b) traffic flow and (c) traffic occupancy over time period 17:45 ~ 18:00 p.m. in different weekdays. Red rectangles indicate the anomalies detected by (Vallis et al., 2014); blue triangles indicate those detected by compression-based method.

Figure 4-1 The ratio of the most frequent visited locations over all locations for 5 individuals over different hour periods; (b) The ratio of the top 10 most frequent visited locations over all locations for the same 5 individuals at 18:00 p.m. in different days.

Figure 4-2 Tweet locations of 5 individuals at 18:00 p.m. in different weekdays and the spatial scale of the most frequent places are labeled with a dashed circle and the radius in km.

Figure 4-3 3 Movements from 17:00 p.m. to 18:00 p.m. in different days: The triangle point is the origin while the dots are destinations.

Figure 4-4 The percentile values of the distances between all locations of two travelers.

Figure 4-5 The frequency of the distances between all locations of two travelers.

Figure 4-6 the distribution of visiting times of all locations. The zero value of x-axis is fixed at the most frequent visited places and the x-axis values indicate the Euclidean
distance of all locations from it. Positive values indicate the locations are on its east while the negative ones on the west. The y-axis values are normalized probability for the number of visiting times; (b) the heat map of visiting times of locations for the traveler. The location of plot (a) is circled.

Figure 4-7 location changes from (a) 18:00 p.m. to (b) 19:00 p.m. and the corresponding tweets: (1): “i hate this haircut”; (2): “ive never seen this movie that they playing on bet”; (3): “xisthatnigga marchmadness this is marchmadness this happening to duke makes this so great”.

Figure 5-1 (a) histogram of the displacements over 1-hour time window of all tweet users in this chapter; the time distribution of the displacements (b) lower than 100-m and (c) higher than 100,000-m; the day-of-week distribution of the displacements (d) lower than 100-m and (e) higher than 100,000-m.

Figure 5-2 Probability density value $P_{SiW}$ of Twitter displacement with two conditioned values: $C=100$-m and $C=2000$-m.

Figure 5-3 Distribution of Twitter displacement percentiles of (a) all tweet users over three different time windows; (b) each tweet user over time window [0, 3600s]; (c) each tweet user over time window [3600s, 7200s]; (d) each tweet user over time window [7200s, 10800s].

Figure 5-4 Geographic distribution of the origins and destinations of the long-distance Twitter displacements of two tweet users. The function $Q()$ calculates the 95th percentile in both plots and the time window is set to be [0, 3600s].

Figure 5-5 Steps of tokenization, stop-word filtering and stemming on a tweet.
Figure 5-6 The long-distance displacement ratio of time window [0, 3600s] to [3600s, 7200s] and [7200s, 10800s] and its corresponding size with different percentile calculator $Q(.)$ ................................................................. 129

Figure 5-7 Long-distance displacements (a) from 5/1/2014 to 5/15/2014, displacements on the same day are in the same color. (b) longer than 3000 m in the whole year. The positions of the start and end are slightly adjusted when they overlap each other in both plots................................................................. 131

Figure 5-8 The counts of daily tweets that contain “Spring Hill”, “Greensboro”, “Tysons Corner”, “McLean”, “Wiehle-Reston East”, “Silverline”......................... 133

Figure 5-9 the locations where the “Silver Line” displacements end (destination) from July 26 to July 28. ................................................................. 134

Figure 5-10 the ratios of CDPs where the “Silver Line” displacements start (origin) from July 26 to July 28. ................................................................. 135

Figure 6-1 Steps of token filtering and stemming ........................................ 140

Figure 6-2 Correlations between the manual label and the individual stemmed tokens ............................................................................................................. 143

Figure 6-3 Regression results with selecting individual tokens under different thresholds of correlation coefficient $\phi$............................................................. 145

Figure 6-4 Regression results with paired tokens under different values of confidence where support is equal to 0.01 ............................................................. 148

Figure 6-5 Comparisons of accuracy and precision between SVMs and sLDA........ 152

Figure 6-6 AIC values for different number of clusters .................................. 156

Figure 6-7 15 different clustered centers of traffic signatures ..................... 156
Figure 6-8 Comparisons between clustered centers and the original traffic flow and occupancy data in two sample detectors

Figure 6-9 Comparisons of accuracy and precision with and without traffic features

Figure 6-10 Distributions of (a) time and (b) space difference between the accident-related tweets and the accident records by VDOT.

Figure 7-1 AIC values for different $k$

Figure 7-2 12 different clustered centers of traffic signatures

Figure 7-3 Distribution of traffic occupancy in two clusters at 7:00 p.m.

Figure 7-4 Percentage of traffic-justified Twitter concentrations under different threshold $ptraffic0$

Figure 7-5 The traffic surge probabilities at different time-of-day that are related to Twitter concentration (1) and (2), listed in Table 7-3. Figure (a) and (b) represent two detectors associated with Twitter concentration (1), Figure (c) and (d) another two detectors with Twitter concentration (2). The dash line indicates when the tweet is blogged.

Figure 8-1 (a) The frequencies of tweet posting time and (b) the frequencies of tweet locations from Jan.1 to May.31 2016 throughout the geography of Ann Arbor region; (c) Month distribution of the same tweet contents from different users.

Figure 8-2 (a) The frequencies of all trip arrival time (b) Trip trajectories of two Connected Vehicles; (c) All trip ends.

Figure 8-3 Flow chart of our study.

Figure 8-4 land use distribution of trip ends by setting space range to be 200 m
LIST OF TABLES

Table 1-1 Tweet samples describing the general traffic information, general traffic incident and road accident, respectively ................................................................. 13

Table 1-2 Data summary .......................................................................................... 15

Table 2-1 Basic statistics of some well-known social media ................................. 27

Table 2-2 Focuses of social media and their representatives .............................. 28

Table 2-3 Part of recent transportation applications on social media ................. 32

Table 3-1 Feature table employed in the regional pattern identification study .......... 58

Table 3-2 An illustrative example of database table and pattern table ............... 59

Table 3-3 Anomaly degrees of spatial traffic flow patterns in 3 sub-regions in different time periods on weekday February 2nd, 2014 ................................................................. 71

Table 3-4 Anomaly degrees of temporal traffic flow patterns in 9 sub-regions in different time periods on the same weekday ................................................................. 75

Table 4-1 Basic statistics for location features of nearly 200 unique tweet users ..... 90

Table 4-2 Part of the keyword and the corresponding social activities .............. 108

Table 5-1 Examples of tokens and their corresponding public activities .......... 125

Table 5-2 Token rankings in two exemplary displacements of a tweet user over time window [0, 3600s] ......................................................................................... 127

Table 5-3 The topic word interpretations of the long-distance displacements (a) Long-distance displacements from 5/1/2014 to 5/15/2014; (b) Long-distance displacements longer than 3000 m in the whole year ................................................................. 132

Table 6-1 Accident-related words ......................................................................... 138

Table 6-2 Paired tokens selected by the Apriori algorithm ............................... 147
Table 6-3 Lists of top 30 topic words (tokens) of sLDA models on tweets .......... 151
Table 7-1 6 keywords and related social events in July 2014.............................. 169
Table 7-2 Transportation lexicon........................................................................ 169
Table 7-3 The keywords, Twitter concentrations and public events corresponding to Figure 7-5................................................................................................................. 181
Table 8-1 Sample tweets with clear trip-related information ......................... 185
Table 8-2 (a) Statistics of tweets around the trip ends and those with clear indications of locations and (b) Quartile values of distance between the tweet locations and the real geographic locations ........................................................................................................ 192
Table 8-3 (a) Statistics of trip destinations with tweet locations; (b) The most frequently visited tweet locations around trip ends ........................................... 193
Table 8-4 Feature ranks of tweet locations around a specific trip location .......... 195
Table 8-5 Comparisons of trip end labels between tweet categories and the previous land use categories ........................................................................................................ 197
ABSTRACT

With the rapid development of communication and detecting technologies and increasing coverage of sensors, large amounts of loop detector data have been collected on a daily basis for transportation studies. The loop detector data can measure the traffic performance, inform the researchers about the status of traffic operation, and finally support the decision-making and improve the traffic in a road network. Besides the traditional traffic-related data such as traffic flow, occupancy, speed, etc., newly emerged data sources, such as social media data, land use information, etc., can find useful traffic-related knowledge in a different perspective and even complement the deficiencies of the traditional studies by crowdsourcing different data sources. Integrating crowdsourced data into transportation study requires the new techniques and methodologies from the fields such as computer science, geography, social network, etc. and the results can possibly provide insights of advanced traveler information. Through several integrative approaches, this dissertation proves the promises of using traditional loop detector data, social media data and other geographic information for advanced traveler information. Our attempts are detailed in the following order:

The first case study is about the traffic flow pattern identification, which is an important component for traffic operations and control. To reveal the characteristics of regional traffic flow patterns in large road networks, we employ the dictionary-based compression theory to identify the features of both spatial and temporal patterns by analyzing the multi-dimensional traffic-related data. An anomaly index is derived to quantify the network traffic in both spatial and temporal perspectives. Both pattern identifications are conducted in three different geographic levels: detector, intersection, and sub-region. From different
geographic levels, this dissertation finds several important features of traffic flow patterns, including the geographic distribution of traffic flow patterns, pattern shifts at different times-of-day, pattern fluctuations over different days, etc. Both spatial and temporal traffic flow patterns defined in this study can jointly characterize pattern changes and provide a good performance measure of traffic operations and management. The proposed method is further implemented in a case study for the impact of a newly constructed subway line. The before-and-after study identifies the major changes of surrounding road traffic near the subway stations. It is found that new metro stations attract more commute traffic in weekdays as well as entertaining traffic during weekends.

The second study reveals the characteristics of travel behavior using high-resolution Twitter data through a series of empirical studies and further explains the abnormal movements extracted from the tweet trajectories. First, we explore the characteristics of individual travel behavior especially the location geo-distribution, movement scale and the clustering features of both the directed and undirected travel. Second, we propose a geo-mobility clustering method that groups the tweet locations driven by the same travel motivation. This clustering method captures the clustering features of a traveler’s hourly locations and detects the abnormal travel behavior. Third, the tweet posts are examined to identify the social activities behind these abnormal movements. The results of our algorithm show that 46.2% of the abnormal movements can be tied with social activities by the keywords of the tweets.

The third study extracts the human mobility patterns from social media and decodes the travel motivations behind the long-distance tweet displacements. We first unveil the characteristics of the tweet displacements and compare the time-of-day features, step-size
features of displacements between social media and other data sources. Second, we define
the human mobility patterns within different time windows and find the motivations that
drive the long-distance displacements. Third, two applications show the great promises of
our studies: we can either unveil the individual mobility patterns or study the travel demand
in a small region given specific social events.

The fourth study combines the loop-detector data and the social media data, to detect
on-site traffic accidents. In previous studies, it remains unknown how effective the social
media based detection methods is as compared with the traditional loop-detector-based
method. In this study, we first explore the features of keywords and their association rules
inherent in the accident-related tweets and explore the potentials of using tweets for
accident detection. Combining the traffic flow and occupancy data, the prediction results
show that some tweets not only respond to traffic accident much more quickly than
traditional methods but also find some undocumented accidents which make up for the
deficiencies of official accident records. Further, the limitations and disadvantages are also
discussed which provide insights in utilizing social media data to assist accurate on-site
traffic accident detection.

The fifth study employs both the traffic data and social media data to study the social
events and their impacts on traffic operations. In this study, the tweets, which are blogged
and shared by the broad masses of people, may be associated with some major social
activities. These tweets are called “Twitter concentrations”. The public activities behind
Twitter concentrations potentially pose more pressure on traffic network and cause traffic
surge within a specified time and location. However, it still remains unknown how closely
the Twitter concentration and traffic surge are correlated with each other. Our study fuses
a set of tweets and traffic data collected during the whole year of 2014 in North Virginia Region, and mainly investigates the correlation between Twitter concentration and traffic surge in July. The results show the promise and effectiveness of our proposed methods and even provide insights in the causality of the non-recurrent traffic surge.

The sixth study proves the credibility and effectiveness of social media in labeling the activity location (trip purpose) for the trip ends. In this study, there are more than 80,000 geo-tagged tweets collected in the area of City of Ann Arbor from January 1 to May 31, 2016. We extract the potential locations embedded in tweets by using selected preposition words: “at”, “in”, “on”, etc. Through manual examination, we find nearly 200 locations that can be verified by the real places. These “tweet locations” can be further categorized into more than 100 location categories by Google Place Types. Our examinations show that for a tweeted trip end, an average of 6 tweets can be found within 50 meters and 30 within 200 meters. Through our ranking algorithm, we show the promises of tweets in inferring the trip purposes. Compared with land use data, the tweets data prove to be a more informative resource in trip purpose inference.
CHAPTER 1 INTRODUCTION

1.1 Research background and motivations

1.1.1 Research background

With the revolution of information, communication, sensing technology, the traffic operations and control meet ever greater challenges and problems. The current traffic problems such as traffic congestion, route navigation, traffic accident, abnormal travel behavior, trip purpose inference, etc. evolve from the increasing transportation demand and the rising of social communication which can be manifested by the ever-increasing prevalence of vehicles. Current intelligent transportation system (ITS) requires the rapid response towards the emergencies, good capability of road facilities, great flexibility of traffic operations and an effective road knowledge feedback system prepared for advanced traveler information. The advanced traveler information for both the commuters and traffic operators may include:

(1) Traffic performance information over different time-of-day.

Studies on traffic flow patterns within a scale of road facilities have aroused increasing attentions in recent years. The traffic flow patterns can be taken as those characteristics of vehicle groups passing a point or a short segment during a specified span or traveling over longer sections of highway (Lan et al., 2008). The spatiotemporal features of traffic flow, occupancy, and speed over different time scales can provide insights into traffic operation and control, urban planning, incident management, etc. The time-of-day and day-of-week features of traffic flow or occupancy reveal the fluctuations of the jammed conditions and traffic operations of road links, intersections or even networks and can serve different research purposes. For instance, White et al., (2007) focused on the impact of the daily
visitor transportation on the public infrastructures and suggested approaches to improve the environmental sustainability of national parks; Ramaswami et al. (1996) studied the network traffic patterns to design the logical topology and the routing algorithm so as to minimize the network congestion; Cassidy et al. (1999) investigated the characteristics of freeway traffic flow patterns and their findings have practical implications for freeway traffic planning and management; Lee et al.(2014) proposed a method to identify the congestion patterns from the big traffic data and predicted when the road congestion events will dissipate.

(2) Travel behavior and human mobility patterns over a geographic scale.

Researchers traditionally center the studies of travel behavior (or human mobility) on trips, including the traffic patterns of trips between origins and destinations (Alexander et al., 2015; Beiró et al., 2016; Simini et al., 2012), the trip schedules such as the departure time or trip periodicity (Marguta and Parisi, 2015; Merler and Ajelli, 2010), and the travel mode selection of trips (Khattak and Rodriguez, 2005). In these studies, the trips are characterized by movements in different spatial ranges as well as at different time perspectives, thus serving different purposes. Long-distance trips may be infrequent including interstate or intercity travel. As compared, the short-distance trips are more common and regular including daily commuting, grocery shopping, or even purposeless wandering etc. and these are the focus of our study. Inferring the motivations of the directed travel and unveiling the human mobility patterns may provide insights to the traffic management and control.

(3) Traffic accident information and other traffic-related activities
The traffic accident is considered as one of the most important urban problems worldwide and may break down the traffic flow and disturb the traffic operations. Major traffic accidents can sometimes cause irreparable damages, injuries, and even fatalities. National Highway Traffic Safety Administration (NHTSA), which publishes yearly reports on traffic safety facts, states that since 1988 more than 5,000,000 car crashes occur in the States each year and about 30% of them bring fatalities and injuries (NHTSA, 2015). After years of research, it has been widely accepted that significant reductions of accident impact can be achieved through effective detection methods and corresponding management strategies. As an essential component of traffic incident management, accurate and fast detection of traffic accidents are critical to modern transportation management (He et al., 2013). Besides the traffic accident, other events that arouse social attentions and possibly bring about travel demand are also worth studying.

(4) Trip information for travelers

Inferring Individual’ activity and trip purposes is critical for transportation planning and travel behavior analysis. Detailed trip information including time, trajectories, and origin and destination locations reveal the trip purposes behind these trips and preferences of the commuters. In the last few decades, the studies of trip purpose (activity location) inference have consistently improved both accuracy and completeness with new proposed methodologies and more types of data involved. A once-straightforward and prevailing method is to utilize the paper travel diaries or phone call surveys to record the trip information as well as the trip purposes (Stopher and Wilmot, 2000). The drawbacks of these methods are as apparent as their merits as it is difficult for respondents to exactly recall the trip-related information including departure time, origins and destinations, etc.
Also, because of the burden of taking these detailed notes all days, the chances of non-response and postponing are relatively high (Bohte and Maat, 2009).

Traditional methods cope with the traffic operation issues by field data within a signalized intersection or a corridor but seldom expand to an entire road network. Problems emerge in most traffic related studies and the knowledge findings and data mining from “Big Data” have become more and more popular. Also, the increasing demand for more informative traffic knowledge such as the trip purpose, better emergency response, etc. In future studies for intelligent transportation, crowdsourcing data from different sources can be of great use and may be the only solution to a better traveler information system. In our study, we combine the traditional traffic data and the social media data to find effective transportation information.

1.1.2 Motivations

The explosion of traffic data in recent years has aroused increasing attentions and amounted to the best interests of intelligent transportation management and control. Loop-detector data including the traffic flow and occupancy, as well as the social media data such as Twitter, can be of great use in mining useful knowledge of highway traffic operations and in turn provide insights in alleviating traffic congestions, detecting traffic accidents, extracting the travel patterns etc.

In recent years, transportation researchers have experienced the benefits of increasing coverages of loop detectors in monitoring the traffic operations, especially in the large metropolitan areas. Our studies show in urban road networks of North Virginia counties: Loudoun, Fairfax and Prince William, there are more than 1500 signalized intersection in monitor and more than 15000 loop detectors collecting traffic flow, occupancy and other
traffic operational data. The traffic data are collected in such great size that the yearly data entries in urban road networks in the Washington D.C. area can amount to hundreds of billion. Also, the social media data can be of even bigger size due to the great coverage of smartphones and the induced needs for social networks. The yearly tweet contexts in New York Metropolitan Area, for instance, can be of tens of gigabytes. The “big data” features of both the traffic data and social media require the best knowledge of data preprocessing, data mining, etc. to explore and discover the hidden patterns and knowledge in delivering the transportation purposes.

At the same time, applications of employing “big data” for transportation analytics are promising which may include the traffic pattern identification, traffic accident detection, travel time evaluation, travel behavior extraction, etc. The “big data” feature can possibly rule out the noises inherent in the traffic operations and even compensate the deficiencies of the traditional methods. Before any fruitful ideas and thoughtful conclusions, one need to fully comprehend the data components and structures and also find proper data prepossessing methods and data mining models.

1.2 Data description

1.2.1 Loop detector data

Our study builds on massive datasets collected by traffic loop detectors in the urban network of Northern Virginia (NOVA), located to the northwest bank of Potomac River and adjacent to the District of Columbia. The land area is more than 2500 sq. km including 3 major counties in Northern Virginia involved in our study: Loudoun County, Fairfax County and Prince William County. The area is further manually divided into more than
21 sub-regions by the Virginia Department of Transportation (VDOT) for traffic operation purposes, shown in Figure 1-1. Furthermore, the area holds nearly 2 million which is about one-quarter of the entire Virginia State and has long been known for its heavy traffic (Cervero, 1994).

We only focus on urban streets, including arterials, collectors and local roads, whereas freeways are not included in the scope of this study. More than 15000 traffic detectors are located on the links of over 1200 signalized intersections. All the detectors are lane-based stop-bar detectors, as shown in Figure 1-1(b). Therefore, the information collected by the detectors not only describe the traffic conditions on certain links but also those of the signalized intersections. The study period contains 8 months from January 1, 2014 to August 31, 2014. The raw data collected include the traffic volume and occupancy in 15 minutes. Both traffic flow and occupancy are widely accepted in studying features of road networks or their impacts on daily traffic, since they can indicate other traffic flow characteristics (e.g. headway, speed, etc.) (Guardiola et al., 2014; Kerner and Rehborn, 1996; Kim et al., 2004). It is worth mentioning that some of the loop detectors may experience different kinds of malfunctions, so data is not completely trustworthy. The entire dataset has been carefully examined, and malfunctioning detectors are excluded from this dissertation.
1.2.2 Social media data

The social media data is a newly emerged data source that can be employed to complement the traditional studies. Some of the popular social media websites in North America include Facebook, Wechat, WhatsApp, Tumblr, WAZE, Instagram, Twitter, Baidu Tieba, Pinterest, LinkedIn, Gab, Google+, Line, YouTube, Viber and Snapchat. We mainly employ the Twitter for our study.

The tweet data were collected through Twitter Streaming API with geo-location filter. Filtering by the coordinates, we extracted tweets posted from both NOVA and Ann Arbor.
region. There are more than 584,000 tweets in a full year from January 2014 to December 2014 in NOVA and more than 50,000 tweets from January to May in 2016. The study areas are shown in Figure 1-2.

![Study area for tweets in (a) NOVA and (b) City of Ann Arbor](image)

Figure 1-2 Study area for tweets in (a) NOVA and (b) City of Ann Arbor

Each tweet posts are coupled with specific date, time and location information. The tweets are the reflection of what people are interested at the specific time and location. Thus, they can justify the traffic accident if the text content has a clear expression of it. The location information is the paired latitude and longitude where the tweets are posted. The resolution of the GPS location can be as high as 100 meters.

Our empirical examinations find there are tweets reporting the traffic-related information. As shown in Table 1-2, even though not all tweet users prefer to give a clear description of the traffic information, there are still some providing evidence for the traffic incident, traffic congestion, etc. In this dissertation, we only study the tweets which have
explicit indications of traffic information and further examine the credibility of these tweets for verification purposes.

Table 1-1 Tweet samples describing the general traffic information, general traffic incident and road accident, respectively

| General information | “I am waiting at the silver line, exciting” |
| General incident | “Always hate the signals ahead of the hip-hop, making me sick” |
| General incident | “standstill for 1 hour, there must be accidents in front” |
| General incident | “this is typical NOVA traffic, what a bad day” |
| Traffic accident | “major accident next to the sunoco near the parkway a car got flipped over” |
| Traffic accident | “the worst car accident possible just happened in front of me” |

1.2.3 Traffic accident data

The traffic accident log was collected from Virginia Department of Transportation (VDOT). The traffic accident log is an accident database of the police reports recording the historical accidents in NOVA in the past few years. There are about 52,496 accidents happen in this study area throughout Year 2014. Each accident is paired with detailed information of latitude, longitude, date, time and corresponding incident description. Such data are taken as the ground truth in proposed classification model to reveal the coverage and effectiveness of tweets in accident detection. The traffic accident data gives a clear description and categorization of the traffic incidents and the classification covers vehicle on fire, crash, tailgating, etc.
1.2.4 Land use data and Google Place Type data

The land use data is acquired online in the Ann Arbor Geographic Information System (2015). The land use data covers all the land information in the region of City of Ann Arbor. It can give a rough category of the facilities. Six major categories are included: “Mixed Use”, “Public /Quasi-Public /Institutional /Organ worship”, “Transportation /Communication /Utilities”, “Office”, “Commercial”, “Recreation” and “Residential”.

By Google Places API, we also obtain the Google Place Types which is assigned to the facilities. The Google Place Type is a category data available online for each commercial, non-profit organization, school, etc. There are certain features of the data:

- First, the data gives an ever-detailed classification type for the facilities. For instance, it will distinguish the “Chinese restaurant” and “Mexican restaurant”. This cannot be achieved by land use data
- Second, the land scale of the Google Place Type is much smaller than the traditional land use data, which means it can give a higher resolution result.

1.2.5 Trip data of connected vehicle

We introduce into our study the Safety Pilot Model Deployment (SPMD) data which is released by the U.S. Department of Transportation’s (USDOT) Real-Time Data Capture and Management program (2015). Taw trip data are available from a recently released CV dataset on Federal Highway Administration (FHWA) research data exchange (RDE) website. The dataset were collected by the safety pilot program, which is hosted by University of Michigan Transportation Research Institution (UMTRI) in City of Ann Arbor.

Our main interests lie in the dataset of Basic safety message (BSM). The BSM is a kind of ‘heartbeat’ message which transmits messages frequently (usually at approximately
10HZ, the same frequency of data used in this thesis). A BSM includes two parts; one part is a binary large object (blob) that includes every BSM. It consists of fundamental data elements that describe a vehicle’s position (latitude, longitude, and elevation) and motion (heading, speed and acceleration). The other part of BSM data contains an extension of vehicle safety information (path history, path prediction and event flags) and pertains to the status of a vehicle’s components (lights, wipers, and brakes).

The trip data were also collected in the region of City of Ann Arbor. It gives us the GPS records in a time interval of millisecond which is accurate enough to capture the vehicle trajectory.

Table 1-3 summarizes and compares the features of all the data that are employed in this dissertation.

Table 1-2 Data summary

<table>
<thead>
<tr>
<th></th>
<th>Traffic data</th>
<th>Twitter</th>
<th>Accident records</th>
<th>Land use</th>
<th>Connected-Vehicle data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured or unstructured</td>
<td>Structured</td>
<td>unstructured</td>
<td>Structured</td>
<td>Structured</td>
<td>Structured</td>
</tr>
<tr>
<td>Size</td>
<td>&gt;20G</td>
<td>&gt;15G</td>
<td>&lt;1G</td>
<td>&lt;1G</td>
<td>&gt;20G</td>
</tr>
<tr>
<td>Area</td>
<td>NOVA</td>
<td>NOVA, Ann Arbor</td>
<td>NOVA</td>
<td>Ann Arbor</td>
<td>Ann Arbor</td>
</tr>
<tr>
<td>Types</td>
<td>Continuous data</td>
<td>Text data</td>
<td>Categorized data</td>
<td>Categorized data</td>
<td>Continuous data</td>
</tr>
</tbody>
</table>

1.4 Contributions

The contributions of this dissertation lie in our systematic studies on employing the crowdsourcing data for advanced traveler information:

(1) Identifying the regional traffic patterns with “Big Data”
The compression theory is proposed to effectively interpret the large collections of multi-dimensional traffic data and anomaly index can quantify the performance of the traffic operations from both the spatial and temporal perspectives. We can reveal the geographic distribution features, time-of-day features of traffic flow patterns through spatial and temporal traffic flow pattern identifications for a large rural road network.

The physical meanings of these anomalies and the comparison results with other methods show the effectiveness and uniqueness of these method. They can be further utilized for the knowledge finding with higher-dimensional traffic-related data.

(2) Decoding the individual travel behavior

We explore the characteristics of individual travel behavior especially the location geo-distribution, movement scale and the clustering features of undirected travel. A geo-mobility based method is proposed to capture the clustering features of traveler’s hourly locations and detect the abnormal travel behavior.

Furthermore, for those abnormal directed travel, we identify the corresponding trip purposes by analyzing the corresponding tweets. Our results show that tweets can be viable data sources to find the travel motivations of the people’s travel, especially those with abnormal travelling patterns.

(3) Unveiling the human mobility patterns and exemplifying the applications

We unveil the human mobility patterns based on the Twitter trajectory data at the first attempt. The human mobility patterns are first defined as the combination of Twitter displacement, Twitter origins and destinations, starting and ending time. The levy flight features of the Twitter displacements are unveiled and compared with previous studies which even show the resolution advantages of Twitter.
At the same time, two applications of the human mobility pattern study are detailed: unveiling the individual travel motivations, especially those long-distance ones and travel patterns in a region towards a special social event.

(4) Detecting traffic accident with social media and traffic data

In addition to separately analyze the words, we reveal the association rules between words in each tweet and includes the association features in our SVMs model for a more accurate accident detection. By employing the SVM model, we study the effectiveness of social media in detecting the traffic accident. Both the individual and paired token features are extracted and regression results show that paired tokens can detect the traffic accidents with better efficiency. In addition, combining with the traffic data will bring no good to the regression results.

We also show the advantages and disadvantages of Twitter as compared to traditional methods: comparing with the official accident records, we unveil that Twitter can report some accidents that are not enlisted in the official records and some accident-related tweets are much faster than the police response.

(5) Detecting the Twitter concentration and its relationship with the traffic surge

First, a probability index is proposed to quantify the level of detector-based traffic surge in a large-scale road network. Second, an effective detection method is proposed to extract, filter and classify the traffic-related Twitter concentrations from a total collection of tweet posts. Third, we develop a methodology to evaluate the correlation between a specified tweet post and its surrounding traffic.
Our results show that the connections between the Twitter concentrations and the public activities. The booms of the tweets that are concerned with a specific topic in an area may indicate a traffic surge in the same place.

(6) Inferring the trip purpose from social media

We study the trip ends of Connected Vehicle and try decode the trip purpose of the trip based on the destination places. The tweets around trip ends show the great promise in giving detailed information of the trip purposes. We found certain advantages: first, the social media reflect the social awareness of the people nearby and can be good representative of the trip purposes. Second, a rankings algorithm is proposed to produce more detailed information and comparison results show that the social media can give more detailed trip information than traditional land use data.

1.5 Dissertation organization

Our research attentions in the dissertation are separately introduced in the following 8 chapters.
In Chapter 2, we conduct a literature review of current studies, especially the deficiencies, which we will respond to in our dissertation. Then we detail our study by first separating the study of loop detector data and social media data, and then combining the two: In Chapter 3, we focus on the loop detector data and conduct the regional pattern identification with the multi-dimensional traffic data. In Chapter 4 and 5, we focus on the social media data and try deriving the knowledge of travel behavior and human mobility patterns from the tweets. In Chapter 6 and 7, we combine the loop detector data and social
media data for the traffic accident identification and Twitter concentration exploration. In Chapter 8, we combine the social media data, land use data, Google place type data and the Connected Vehicle data to infer the trip purpose from tweets.
CHAPTER 2 LITERATURE REVIEW

2.1 Regional traffic flow pattern identification

2.1.1 Gaps in large-scale traffic pattern identification

Of all these studies related to the traffic flow pattern identification, despite their different emphasis and purposes, the traffic pattern identification usually performs as an initial step in the entire research process in transportation management, and there is still research gaps to explore further.

First, the measures for traffic flow patterns are not limited to traffic flow measurement. The metrics to describe the patterns usually vary from traffic flow (Cassidy and Bertini, 1999; Shen and Zhang, 2009; Zhang et al., 2001), density (Cassidy and Mauch, 2001; Treiber and Kesting, 2012), speed (Banaei-Kashani et al., 2011), etc. This is mainly due the feature of each metric being unique and the researchers usually have different purposes that their problems intentionally shape the definition of the traffic flow patterns. Besides traffic-related metrics, other information such as time and link locations may further enrich and clarify the connotations of pattern identification. For example, studies show that given the combination of direction, connectivity and locality of a road segment (and without having access to the actual traffic flow in the segment), one can distinctively determine the corresponding traffic signature (pattern) of a road segment with high probability (Banaei-Kashani et al., 2011; Zhang et al., 2016a). The features of the traffic flow patterns should be further explored combining both traffic, time and location information.

Second, traditional studies focused mainly on traffic flow patterns within intersections (Teodorovic et al., 2001) or corridors (Lan et al., 2008; Schoenhof and Helbing, 2007). These may not be sufficient to explain the pattern changes in a large-scale road network.
with hundreds of intersections. In recent years, the size of traffic data booms both in time and space, and the large-scale floating and fixed sensors are extensively utilized to collect the traffic data. The studies of pattern identification can expand from links to a district or even to a county. This expansion brings more difficulties and has aroused increasing attentions. There have been efforts in reducing the dimensionality of the traffic flow pattern problems, such as functional data analysis (Guardiola et al., 2014), principle component analysis (Jin et al., 2008). How to archive and summarize massive historical data effectively and extract meaningful traffic flow patterns from accumulated data to support decision making has become a significant challenge, considering the huge size of the dataset (Xu et al., 2013).

Third, traditional studies of traffic patterns drew their conclusions from different spatial and temporal perspectives including the day-to-day features (Hu and Mahmassani, 1997) or time-of-day features (Anbaroglu et al., 2014). The temporal traffic patterns obtained from traffic time series studies can reveal the pattern features in a different perspective (Li et al., 2015). Of all these studies, there is a pressing need to develop a systematic approach for traffic flow pattern identification and anomaly detection that consider both different spatial and temporal perspectives. The temporal features of the traffic flow pattern, e.g. at an intersection, should be unveiled by comparing with its surrounding traffic as well as its historical ones. Instead of combining “time” as a traffic-related variable, we proceed with a temporal perspective to study the changes of traffic flow patterns over time.

2.1.2 Pattern identification and anomaly detection

Literatures of anomaly detection in computer science partially or even wholly originates from the study of pattern classification or categorization. The importance of
categorization was stressed together with the proposal to discriminate and identify (perceptual) stimuli as unique items (Bruner and Austin, 1986; Nosofsky, 1986) in 1956. The patterns in early literatures are classified first and an unclassified sample points are assigned to the classification of the nearest set of previously classified points by “the nearest neighbor decision rule” (Cover and Hart, 1967). Then, four problems of categorization are addressed and furthered in later literatures: First, classification was further broadened by the definition of “Ad hoc categories” (Barsalou, 1983) which are created instantaneously to fulfill certain purposes; Second, the lower and upper bounds of the categories are studied and defined (Barsalou, 1983; Shamos and Hoey, 1975; Valiant, 1984) and there is no all-embracing methodology; Third, many mathematical methods or models are explored to describe and quantify pattern classification such as nearest neighbor methods (Roussopoulos et al., 1995; Zahn, 1971), smallest-circle problem (Elzinga and Hearn, 1972), Euclidean minimum spanning tree (Chang, 1974), voronoi diagram (Shamos and Hoey, 1975), hidden Markov models (Smyth, 1997) etc. Fourth, algorithms are impelled to a constant improvement to gain more effective and efficient computation because of the nature of high dimensionality and large size of data sets (Aggarwal et al., 1999; Agrawal and Srikant, 1994; Knox and Ng, 1998). The detrimental effects of the curse of high dimensionality have been studied in great detail and substantially diminished (Aggarwal, 2001). The current state-of-the-art technique to indexing high dimensional data is to first reduce the dimensionality of the data using Principal Component Analysis and then indexing the reduced dimensionality space using a multidimensional index structure (Chakrabarti and Mehrotra, 2000). Historic view of the literatures in pattern
analysis serve well as references to transportation fields and a better understanding of categorization contributes to traffic pattern identification.

On this basis, study of anomaly detection, which has been taken as a side product of clustering algorithm (Aggarwal et al., 1999; Aggarwal and Yu, 2000, 2001), thrives in diverse research areas and application domains such as detecting intrusions against programs (Ghosh et al., 1999; Ghosh et al., 1998; Lane and Brodley, 1997), mining unexpected knowledge (Knorr and Ng, 1999; Knox and Ng, 1998), characterizing the behavior of the system (Lane and Brodley, 1999). Anomaly detection is also interpreted into different ways: In “deviation detections”, by comparing a series of similar data, elements that disturb the series are considered exceptions (Arning et al., 1996). In “novelty detection”, “novelty” is defined as deviations exceeding an allowable variation from the normal data pattern (Dasgupta and Forrest, 1996). In contrast, compression based methods are more straightforward in stressing that the usage of a certain pattern is an important attribute to measure its anomaly degree (Akoglu et al., 2012). As data stored and transferred contains significant redundancy (Welch, 1984), the problems of compression was once approached using “Minimum Description Length principle” (MDL principle) to pull through the difficulty of finding the most interesting frequent item sets (Siebes et al., 2006). In MDL principle, any regularity in the data can be used to compress the data (Grünwald, 2005), so the best set of frequent item sets is the set that compresses the database best (Siebes et al., 2006; Vreeken et al., 2011). To finish this task, a code table is needed to assign code words to categories or patterns and the patterns with short code words are those that have high usage, representing the patterns in the data that can effectively compress the majority of the data points (Akoglu et al., 2012). Thus, the anomaly scores, also referred
as the classification scores, can be calculated according to the length of code words and represent how significantly a data (pattern) deviates from learned normal patterns (Hirose et al., 2009). Theoretically, the least frequent item sets are the rarity parts which can be taken as kinds of anomalies. What to mention is that most of the compression anomaly detection deals with binary or categorical data while the traffic operators usually manage continuous or sequential data. A pre-categorization process for continuous traffic data is necessary.

2.1.3 Dimensionality reductions of traffic data

To cope with the large dimensionality of the traffic-related data, we propose a method based on the compression theory in regional traffic flow pattern identification. Compression-based approaches have been successfully implemented in pattern recognition and anomaly detection in different domains, such as image processing (Akoglu et al., 2012), system query processing (Binnig et al., 2009), etc. Rather than trying to compress the set of frequent items, compression-based approaches compress the database and search for the subset of all frequent item sets that compresses the database best (Siebes et al., 2006). The method can capture the best regularities of the data with as little redundancy as possible and thus can capture the most important patterns in the datasets (Tatti and Vreeken, 2008). Previous research proves that this approach provides accuracy on par with the state-of-the-art in anomaly detection (Smets and Vreeken, 2011). Some even argue that this approach is competitive or superior to many of the state-of-the-art approaches in anomaly and interestingness detection, classification, and clustering with empirical tests on time series, DNA, text, XML, and video datasets (Keogh et al., 2007). Similar to the data fusion which can maximize the utility of the available (traffic) information (Treiber and Kesting, 2013),
the compression method is capable of recognizing the frequent traffic flow patterns through effective interpretations of multi-dimensional data. Based on the compression theory and corresponding algorithm, we will be able to quantify the traffic performance in a given location and detect the abnormal traffic flow patterns from our studied network.

2.2 Review of social media

2.2.1 Social media definitions and features

Social media are computer-mediated technologies that allow individuals, companies, NGOs, governments, and other organizations to view, create and share information, ideas, career interests (Buettner, 2016). Obar et al. (2015) synthesize the definitions in literature and identify the following commonalities among current social media services:

- Social media are interactive Web 2.0 Internet-based applications.
- User-generated content such as text posts or comments, digital photos or videos, as well as data generated through all online interactions, are the lifeblood of the social media organism,
- Users create service-specific profiles for the website or app, that are designed and maintained by the social media organization, and
- Social media facilitate the development of online social networks by connecting a user's profile with those of other individuals and/or groups.

The social media has long been seen in the traditional news, media or websites and has evolved with the transformation of communication technologies. The technical evolution such as RFID, open web standards is laying the groundwork for moving Social Media applications away from desktop PCs and laptops, toward mobile devices (Kaplan and
Haenlein, 2010). The evolution comes with the increasing coverage of smartphone and other mobile devices. The social media convert the traditional one-directional news feed from large media organizations to the individuals, into a bidirectional information sharing where everyone becomes a news center. The well-known social media websites include Facebook, Wechat, Tumblr, Instagram, Twitter, Baidu Tieba, Pinterest, LinkedIn, Google+, YouTube, Viber and Snapchat. Their basic information is shown in Table 2-1. The statistics of monthly active users are acquired from DMR websites (DMR, 2016) and Statista.com (Statista, 2016) and may be an approximate number due to the statistics sources.

Table 2-1 Basic statistics of some well-known social media

<table>
<thead>
<tr>
<th>Social media name</th>
<th>Monthly active users (M)</th>
<th>Release date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>1650 (3/31/2016)</td>
<td>2004</td>
</tr>
<tr>
<td>Wechat</td>
<td>700 (4/31/2016)</td>
<td>2011</td>
</tr>
<tr>
<td>Tumblr</td>
<td>555 (9/1/2016)</td>
<td>2007</td>
</tr>
<tr>
<td>Instagram</td>
<td>500 (6/21/2016)</td>
<td>2010</td>
</tr>
<tr>
<td>Twitter</td>
<td>310 (5/1/2016)</td>
<td>2006</td>
</tr>
<tr>
<td>Baidu Tieba</td>
<td>657 (2/5/2016)</td>
<td>2003</td>
</tr>
<tr>
<td>Pinterest</td>
<td>100 (9/18/2015)&lt;sup&gt;(1)&lt;/sup&gt;</td>
<td>2010</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>106 (4/28/2016)</td>
<td>2002</td>
</tr>
<tr>
<td>Google+</td>
<td>540 (2/1/2016)</td>
<td>2011</td>
</tr>
<tr>
<td>YouTube</td>
<td>1000 (4/21/2014)</td>
<td>2005</td>
</tr>
<tr>
<td>Viber</td>
<td>236 (2/12/2015)</td>
<td>2010</td>
</tr>
<tr>
<td>Snapchat</td>
<td>100 (8/1/2015)&lt;sup&gt;(2)&lt;/sup&gt;</td>
<td>2011</td>
</tr>
</tbody>
</table>

<sup>(1)</sup>This is total number of Pinterest users.
This is the daily active users of Snapchat

Social media technologies take many different forms including blogs, business networks, enterprise social networks, forums, microblogs, photo sharing, products/services review, social bookmarking, social gaming, social networks, video sharing, and virtual worlds (Aichner and Jacob, 2015). Each social media website and tool has its own specialized focus. For instance, Twitter focuses on sharing short message streams; LinkedIn is mainly for business networks and career purposes; Square shares the dining and recreational places; Baidu Tieba focuses on the topics and corresponding discussions; Wikipedia is a knowledge-sharing website; Facebook builds platforms for social networks. Table 2-2 summarizes the focuses of social media and the representative social media websites and tools in North America. The focuses are subject to change and involve with times passing by and new forms of social media are coming forth continuously.

Table 2-2 Focuses of social media and their representatives

<table>
<thead>
<tr>
<th>Focus</th>
<th>Social media website and tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live casting</td>
<td>Actors Access, Backstage, Actors Equity Casting Call, Playbill, SAG Indie, Now Casting, Casting Networks, NYCastings.com, Mandy.com, Craigslist</td>
</tr>
<tr>
<td>Virtual world</td>
<td>OurWorld, Wizard 101, Woozworld, Virtual Families, Second Life, IMVU, Habbo Hotel, Smeet, Meez, SmallWorlds</td>
</tr>
<tr>
<td>Category</td>
<td>Services</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Music</td>
<td>Pandora, Yahoo! Music, Google Play, SoundCloud, Spotify, MySpace, Tunein, Last.fm, iHeart, AllMusic, Jango, Radio, Songza, live365, slacker</td>
</tr>
<tr>
<td>Event sharing</td>
<td>Beadwork, Folio, Forbes, HOW, Dwell, Farm Progress, WineMaker, Entrepreneur, Wired</td>
</tr>
<tr>
<td>Document sharing</td>
<td>ISSUE, SlideShare, Scribd, Box.net, Docstock, Calameo, Zoho, Keep and Share, Free eBooks, 4shared, Author Stream, Wattpad, YUDU, Wuala, Div Share, ADrive</td>
</tr>
<tr>
<td>CRM (customer relationship management)</td>
<td>Salesforce, Microsoft Dynamics CRM, Oracle Sales Cloud, SugarCRM, Workbooks CRM, Insightly, Nimble, Zoho CRM, NetSuite CRM, Veeva CRM</td>
</tr>
<tr>
<td>Video</td>
<td>YouTube, Vimeo, Yahoo! Screen, DailyMotion, Hulu, Vube, Twitch, LiveLeak, Vine, UStream, Break, TV.com, MetaCafe</td>
</tr>
<tr>
<td>Business relationships</td>
<td>AngelList, Beyond, Black Business Women Online, Data.com Connect, EFactor, Gadball, Gust, LinkedIn, Meetup, Networking for Professionals, Opportunity, PartnerUp, PerfectBusiness, Plaxo, Quibb, Ryze, StartupNation, Upspring, Viadeo</td>
</tr>
<tr>
<td>Dashboards</td>
<td>Toutapp, Lancaster Bingo Company, A flat design dashboard, Yet more flat design, fitbit, Patient records, Sprout Social, Nektar Dashboard,</td>
</tr>
<tr>
<td><strong>General</strong></td>
<td>Facebook, Hi5, myspace.com, renren, bebo, perfspot networking</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Discussion boards and forums</strong></td>
<td>phpBB, Simple Machines Forum, Zetaboards, bbPress, Vanilla Forums, PunBB, fluxBB, PlushForums, Phorum, MyBB, miniBB</td>
</tr>
<tr>
<td><strong>DIY and custom</strong></td>
<td>Ana-White.com, Shanty-2-Chic, Jay’s Custom Creations, ArtofManliness.com, Instructables</td>
</tr>
<tr>
<td><strong>Microblog</strong></td>
<td>Twitter, Friend Feed, Tumblr.com, Plurk, Pinterest, Flattr, Dipity, Yammer, Meetme, Plerb</td>
</tr>
<tr>
<td><strong>(Q&amp;A) Questions and answers</strong></td>
<td>Quora, Mind the Book, Amazon’s Askville, Yahoo! Answers, Stack Overflow, Super User, LinkedIn Answers, Answers.com, Hacker News’ Ask Section, LawPivot</td>
</tr>
<tr>
<td><strong>Social commerce</strong></td>
<td>Pinterest, Shopee, Lyst, Soldsie, Kickstarter</td>
</tr>
<tr>
<td><strong>Pictures</strong></td>
<td>Instagram, Imgur, Flickr, Photobucket, DeviantArt, Shutterfly, TinyPic, WeHeartIt, ImageShack, ImageVenue, SmugMug</td>
</tr>
</tbody>
</table>

### 2.2.2 Social media in transportation application

Despite the different types of social media, the social media has common features which make them useful for transportation applications:
- Social reflection: social media can reflect the social trends of the massive crowd of people and this makes possible the related studies such as traffic accident studies, traffic jam and social event detection, etc.

- Cheap: the social media are mainly the online data that can be obtained through the open API. Now, the most popular API are from the Twitter Inc.

- Multi-topics: as the social media are the reflection of people's social networks. The latent topics in the social media are various and complex and almost all human-related activities can be found from the social media.

Besides these merits, there also exist the challenges to be solved before using social media for research purposes, which can be generalized as:

- Noise: usually, the social media data need to be cleaned and filtered as there are abnormal data and meaningless sample in the data.

- Unstructured: the words, pictures, and forms of social media data are quite different from traditional traffic-data including traffic flow, occupancy, etc. Some of the data need to be categorized and others need to be structured into a database.

- Untrustworthy: some of the information may not be fake and untrustworthy. Sometimes, the information needs to be cross-validated before put into use.

- Vague location: the location information is very important to the transportation study. However, sometimes, the location may be vague due to the GPS errors, etc.

These challenges can usually be countered by introducing the computer science technologies, nature language processing, etc. The related transportation applications have thrived in recent years. Table 2-3 listed parts of the applications on social media and more related studies are under go.
Table 2-3 Part of recent transportation applications on social media

<table>
<thead>
<tr>
<th>Application</th>
<th>Author and year</th>
<th>Social media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel information retrieval</td>
<td>Xiang and Gretzel (2010)</td>
<td>Google search</td>
</tr>
<tr>
<td></td>
<td>Xiang et al. (2010)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Ueno et al. (2012)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Evans-Cowley and Griffin (2012)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Lee et al. (2015)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Lin et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>Social event detection, Traffic incident detection</td>
<td>Sakaki et al. (2010)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Krstajic et al. (2012)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Schulz et al. (2013)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. (2015)</td>
<td>Twitter</td>
</tr>
<tr>
<td>Disaster relief</td>
<td>Gao et al. (2011)</td>
<td>Ushahidi</td>
</tr>
<tr>
<td></td>
<td>Ukkusuri et al. (2014)</td>
<td>Twitter</td>
</tr>
<tr>
<td>Transportation planning, Transportation policy-making, commercial service, Transit management</td>
<td>Evans-Cowley (2012)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Camay et al. (2012)</td>
<td>Facebook, Twitter, Flickr, etc.</td>
</tr>
<tr>
<td></td>
<td>Stambaugh (2013)</td>
<td>Facebook, Twitter, YouTube, etc.</td>
</tr>
<tr>
<td></td>
<td>Gelernter et al. (2013)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Pender et al. (2014)</td>
<td>Facebook, Twitter</td>
</tr>
<tr>
<td></td>
<td>Chan and Schofer (2014)</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>Ni et al. (2015)</td>
<td>Twitter</td>
</tr>
</tbody>
</table>
One can see that most of the social media studies are based on Twitter because Twitter data can provide explicit date, time information and geo-tagged tweets even provide the longitude and latitude location information. The data can be collected in large quantities given the provided Twitter API. There are also other 3-party software platform that can extract data from LinkedIn, Facebook, etc. and these data should match certain search criteria. Thus, they cannot be collected in large quantities and usually designed for specific research purposes.

2.3 Travel behavior and travel motivation identification

2.3.1 Mobility features of travel behavior based on Twitter

Past studies tend to clarify the movements by predefining the scales of zones (Balcan et al., 2009; Lenormand et al., 2014) and in most of these studies, the distance range of the movement in each trip may vary from hundreds of meters to a few miles. This attempt in the travel behavior studies can not only unravel the interesting properties of the underlying mobility patterns (Schneider et al., 2013) but also have some practical meanings such as features of epidemics spreading (Anderson et al., 1992), population diffusion (Petrovskii and Morozov, 2009), social networking (Centola, 2010), etc. As compared to unveiling the trip features of the majority, it seems a bigger challenge to find out reasons that drive individuals to move. Compared to travel behavior studies on the groups, explaining the travel motivations of individuals is more difficult due to the great varieties of human
activities, the unpredictable natures of human mind and the lack of explicit and detailed information.

Previous studies categorized travels into directed travel and undirected travel (Schneider et al., 2013). The directed travel holds the common view that human mobility is a motivation-driven activity. These motivations are general human mobility characteristics that can account for the activities mostly. Aarts et al. (2000) concluded that human travel behavior is sometimes habitual and when behavior is habitual, behavioral travel responses are conditional on the presence of a travel goal. The directed travel is commonly seen, such as travelling to a conference meeting, commuting to workplace, going to school etc. In contrast, undirected travel comes from the idea that the destination is sometimes ancillary to the travel rather than the converse which is usually assumed (Mokhtarian and Salomon, 2001). His studies show that the share of total travel that is completely undirected is presumed relatively small but this travel is not a byproduct of the activity but itself constitutes the travel activity (Mokhtarian and Salomon, 2001). The undirected travel is also common, such as tourists wandering around, joy-riding, strolling in the park etc. in which travelers’ attitudes and personality are more important determinants than objective travel (Ory and Mokhtarian, 2005). The destination locations and the corresponding motivations of undirected travel are usually difficult to predict because under most circumstances they are controlled by undetermined travelers’ minds. As not driven by the special needs and purposes, undirected travel is expected to be shorter than directed travel and the spatial scale of undirected travel should be possibly limited in order that it does not influence the directed travel driven by daily motivations. This is common in scenarios like people will always go back home or hotel after leisure walking.
or jogging near the lake. Thus, the effects of undirected travel on the travel behavior studies are not as significant as that of directed travel and the undirected travel do not give hints to major motivations of directed travel because it results from the desire for travel not specific activities. In sum, it is the directed travel, not the undirected one, which can unveil the travel motivations and decode the travel behavior.

If so, the studies of extracting travel behavior features from on-line tweet posts (Hawelka et al., 2014), cellphone records (Gonzalez et al., 2008), land use information (Wolf et al., 2001), etc., may be greatly influenced by the undirected travel, because the data sampled during the process of undirected travel may disturb the our determination of the true travel motivations. This influence can possibly get worse with the increasing sizes of the predefined study areas. By initially examining the locations obtained from high-resolution geo-tagged Twitter data, we found that some of the movements made by the travelers share similar destinations that, although their geographic coordinates are not exactly the same, fall into the same cluster. As shown in Figure 2-1, the locations in the same cluster are those driven by the same travel motivations while those in different clusters are just the reverse. It needs to be explained whether there are some undirected travels that take place around the certain locations and consequently whether locations in the same cluster, although not exactly in the same place, are driven by the same purpose. If they are both true, one can say that undirected travels involve movements that are as compared much smaller and more unpredictable.
Figure 2-1 546 Twitter locations of a person at 17:00 p.m. in different weekdays. The locations are visually clustered where the size of Cluster 1 is 481, Cluster 2 is 19 and Cluster 3 is 21.

2.2.3 Travel motivation identification

On full considerations of the effects of undirected travel, there are still directed travel, driven by special travel motivations, on a certain day that is quite different from other days. However, the underlying causes to the travel patterns as shown in Figure 2-1 are usually complex due to the fact that there are so many potential influential factors. First, it is well-known that routine activities lead to predictable mobility patterns and our daily mobility (travel) is, in fact, characterized by a deep-rooted regularity despite our deep-rooted desire for change and spontaneity (Song et al., 2010). Other than routine activities, major social events such as running races or grand festivals will inevitably gather thousands or even millions of people (Stange et al., 2011) and thus affects their mobility patterns. Besides, some studies showed that land-use patterns impact the travel modes selection and
consequently affect the mobility patterns. All else equal, residents of neighborhoods with higher levels of density, land-use mix, transit accessibility, and pedestrian friendliness drive less than residents of neighborhoods with lower levels of these characteristics (Handy et al., 2005). In contrast, some studies show the causality between land use and peoples’ preferences can be just reverse: the prevalence of walking and transit use may be caused by self-selection; that is, people who prefer walking or transit may choose neighborhoods that support their predilections (as opposed to neighborhood designs strictly influencing choices) (Ewing and Cervero, 2001). They believe the attitude of the traveler and social influence greatly influence travel behavior (Chen et al., 2015; Xiao and Lo, 2016). The mobility differences are largely explained by attitudes and the effect of the built environment mostly disappears when attitudes and socio-demographic factors have been accounted for (Handy et al., 2005). Some studies present the critical role of people’s social network. A higher number of repeated interactions led to greater number of trips, while a larger social network size led to visiting more locations (Silvis et al., 2006). One can say that the individuals' travel behavior is conditional upon their pursuit of social activities, that is, a key cause of travel behavior is the social dimension represented by social networks (Carrasco et al., 2008). Also, people with higher mobile phone usage have a larger spatial extent regarding their movement patterns (Yuan et al., 2012). Despite the ambiguous cause-and-effect relationships between travelers’ attitude and the external influential factors, one can admit that individuals' communication and social activity patterns can thus be inferred in part from their social networks (Carrasco et al., 2008).
2.2.4 Research gaps and opportunities

Besides above possible assumptions and even quantitative explanations in the human mobility and travel motivation inference studies, there still exist several research gaps and opportunities in explaining travelers’ movements:

First, the resolution of the data in existing studies is still too low to discover detailed traveler information for travels in short distance within a county. For example: the bank notes (Brockmann et al., 2006) are usually used to trace the inter-state or inter-city travel and the corresponding time gaps are usually more than days. Popular data sources from cellphone (Song et al., 2010) which can access a more frequent mobility data but the geo-precision is in kilometers because the average service area of each mobile tower is approximately 3km² (Gonzalez et al., 2008). If the resolution of the data can increase both in time and space, it will be of great use in explaining the travel behavior. Correspondingly, new problems emerge in how we can eliminate the effects of undirected travels which lack specific trip purposes.

Second, some of the methods used only serve well in explaining the human travel under specific circumstances and the results come from the deliberately designed environment instead of instantaneous situations on the scene. For instance, in the studies of Silvis et al. (2006) or Aarts et al. (2000), the experiments or survey access directly the explanations of these human mobility but sometimes is subjective and empirical and usually influenced by the experiment designers. Other methods using GPS tracer (Rhee et al., 2011) suffer the problems of small size of participants and the possibility of losing generality. Same problems also exist in the newly emerging data source of Flicker (Beiró et al., 2016) because of its specific purpose-design for photo sharing.
To make up for the above shortcomings and broaden the horizons of travel behavior studies, data crowdsourcing may be a good option as suggested in related studies such as (Chen et al., 2016; Lin et al., 2013; Wang et al., 2016; Zheng et al., 2016). We employ the geo-tagged Twitter data (i.e. tweets) to explore the features of travel behavior, especially those of abnormal ones. Twitter, the microblogging service which has received increasing attentions in recent years, has been gradually accepted as a user-contributed data source in event detection and summarization such as natural disasters (Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events (Shirky, 2011), etc. Compared to bank notes and cellphone records, geo-tagged tweet data are enabled with the higher geo-resolution with less than 100 meters. The frequency of tweets by a single user can be as high as 100 per day and can build easy access of individual’s hourly trajectory. Posted together with the geo-locations, tweets also serve as a convenient subject to explore the on-site description of travel motivations instead of post-event recall.

there are even challenges to be addressed for travel motivation identification based on tweets: First, the language of tweets is inherently complex, unstructured and ambiguous. Previous studies on text processing employ the support vector machine (D’Andrea et al., 2015; Schulz et al., 2013), natural language processing (Li et al., 2012; Wanichayapong et al., 2011), etc. to automatically process the language for research purposes. However, as the length of a tweet content is sometimes much shorter than traditional news broadcasts and web articles, mining useful information from tweet contents are sometimes more difficult. Second, the travel motivations are usually of great variety and the corresponding public activities are sometimes too difficult to generalize. There are attempts to predefined the urban activities as “Home”, “Work”, “Eating”, “Entertainment”, “Recreation”,

39
“Shopping”, “Social Service” and “Education” (Hasan and Ukkusuri, 2014). Even though, the results still need to be interpreted with caution. How to generalize the labels of the public activities and further employ the travel motivations for the purposes of traffic management and control is a major challenge.

2.3 Traffic accident and social event detection

2.3.1 Traffic accident detection based on social media

Due to the fact that major traffic accidents potentially interrupt the traffic flow, traditional attempts in traffic accident detection (Zhang et al., 2016f) focus mainly on monitoring fluctuations and changes in one or more traffic-related metrics such as the traffic flow, occupancy, speed, etc. Some methods leverage the time-of-day characteristics and geographic features to identify the anomalies that may indicate a traffic accident as traffic measurement features in incident conditions are significantly different from those in normal conditions. For example, Tsai et al. (1979) applied a pattern-recognition approach to improving incident detection algorithms; Sethi et al. (1995) separated incidents by locations and achieved a better detection rates by measuring the speed difference and speed ratio; Samant et al. (2000) developed an effective traffic incident detection algorithm to extract incident-related features from traffic patterns. Jin et al. (2009) detected traffic incidents on the basis of traffic flow-occupancy relationships. With years of dedications within the field, more and more detector-data-based detection algorithms are becoming available, including various regression analysis (Sethi et al., 1995; Yuan and Cheu, 2003), Artificial Neural Network (Khan and Ritchie, 1998), Bayesian-based Network (Abdulhai and Ritchie, 1999; Zhang and Taylor, 2006), Time series algorithms (Teng and Qi, 2003;
Willsky et al., 1980), etc. One of the best known practical algorithms is California algorithm developed by Payne et al. (1978). The California algorithm used freeway traffic flow data for the detection of accidents and other lane blockage incidents that temporarily disrupt traffic flow. Another widely tested algorithm was developed by Hall et al. (1993), known as the McMaster algorithm. Stephanedes and Chassiakos (1993) developed the DELOS 3.3 algorithm, also known as the Minnesota algorithm. It is worth mentioning that besides measurements from loop detectors, the probe vehicle data also proves to be a reliable data source and can be included in the fixed detector and probe vehicle algorithms in studies like (Sethi et al., 1995). Similarly, Amin et al. (2012) proposed to utilize the capability of a GPS receiver to monitor the speed of a vehicle and detect accident based on monitored speed; Park et al. (2015) estimated incident impacts and incident detection by using probe vehicle techniques, etc.

Despite the adaptabilities of above studies, traditional detection methods with only traffic data still meet certain challenges. First, most of the previous research, which utilized the field data to detect the traffic accidents, build on an implicit assumption that the data is reliable. However, detector failures and communication errors are perennial problems in traffic operations. For example, Illinois Department of Transportation (IDOT) in Chicago reported that around 5 percent of their loops (detectors) are inoperative at any given time (Kell et al., 1990). The problem of malfunctioned sensors could cause even more troubles in accident detection in large regions, say, an area with more than 10,000 signalized intersections. Second, the uncertainty nature of traffic patterns and non-recurrent events may undermine the potential of traffic metrics in justifying the traffic accidents. Besides traffic accidents, daily traffic operations may suffer breakdowns by other factors such as
parades, road constructions, running races, etc. Thus, the metrics including the traffic flow and occupancy inherently perform as an indirect support for traffic accidents instead of a direct proof. To address these challenges, there are efforts in applying clustering or classification methodologies such as K-means (Münz et al., 2007) on large data collections to diminish the errors.

Different from data sources from loop detectors or on-road vehicles, Twitter, the microblogging service that has received increasing attentions in recent years, has been gradually accepted as a direct user-contributed information source in event detection. Twitter establishes an online environment where the content is created, consumed, promoted, distributed, discovered or shared for purposes that are primarily related to communities and social activities, rather than functional task-oriented objectives (Gal-Tzur et al., 2014). Thus, in Twitter each tweet acts as a data source of “We Media” and it is possible to retrieve the wide-range information from the broad masses of people in a timely manner. What is more, as more users tweet on mobile devices than on PC (Protalinski, 2012), the corresponding time and location information along with Twitter will be of great use in the detection and broadcasting of social events including earthquakes (Sakaki et al., 2010), bird flus (Aramaki et al., 2011), politic events (Shirky, 2011), etc. The location effectiveness and timeliness features of Twitter can even provide a side-proof in a recent accident detection study that uses the GPS-enabled smartphones (White et al., 2011). As tweets are able to describe what is happening on the scene site and the tweeting locations may be quite near the scene site, the tweet contents are usually the priority in most of the studies through automatic detection of words in tweets especially those that occur disproportionally frequently at the current time (Giridhar et al., 2014). There are also
similar trials in the domain of traffic incident detection. For example, Mai et al. (2013) compared incident records with Twitter messages and proved the potentials for information from Twitter to add context to other traffic measurements as a supplemental data source. Schulz et al. (2013) used microblogs to detect the small-scale incidents; Gal-Tzur et al. (2014) conducted a corridor study on the correlation between tweet and traffic jam. Gu et al. (2016) combined the data sources from Twitter, incident records, HERE, etc. and employed the Naïve-Bayes classification to detect five major incident types; D'Andrea et al. (2015) compared accuracies and precisions of different regression models including Naïve-Bayes, Support Vector Machine, Neural Network, Decision Tree in detecting traffic incidents from Twitter stream. Most of these studies focus on methodologies of automatic information extraction from tweet contents through the techniques of natural language processing (NLP).

Instead of studying all traffic-related incidents, Our study limits the scope in specifying the traffic accident including “collision”, “disabled vehicle” and “vehicle on fire”. Under this purpose, this study employs two major supervised learning models: support vector machines (SVMs) (Cortes and Vapnik, 1995) and the supervised latent Dirichlet allocation (sLDA). From the view of SVMs, a tweet post can be disintegrated into a bag of words and those traffic-related words such as “accident”, “crash”, etc. can be taken as important features in the model. Besides single word features, the correlations among these traffic-related words are also critical because people always describe a topic with a group of words that contents more specific indications than single words. Thus, the association rules between words can possibly increase the prediction precisions of the model and are also fully considered. From the view of sLDA, a tweet post can be disintegrated into a bag of
topics. The proportions of those topics can be approximated by the Dirichlet distribution and even be inferred from word distribution in each topic in tweets. According to Mcauliffe and Blei (2008), the parameters of the topic and word distributions can be inferred by Expectation-Maximization (E-M) algorithm based on the labeled word documents. These words associated with a topic should be further examined. Our contributions can be summarized as: First, in addition to separately analyze the words, this dissertation reveals the association rules between words in each tweet and includes the association features in our SVMs model for a more accurate accident detection; Second, as a first attempt, this study fuses the traffic-related data and the tweet data to further explore the possibility to increase detection accuracy and precision. The role of traffic features in improving the accident detection is discussed further. Third, this dissertation also compares the results of SVM and sLDA. The drawbacks of using topic models in accident detection are revealed. Fourth, by comparing the prediction results of several models with the ground truth from the traffic management log, it is found that the tweets can possibly supplement the current official accident records. The advantages and disadvantages of accident detection based on tweets are also discussed.

There are major challenges to be addressed before the use of tweets in traffic accident detection. First, as compared to events that arouse enormous public attentions such as sporting games, extreme weathers or traditional festivals (Zhang et al., 2016e), the influence of traffic accidents are comparably a “midget”. From our observations, tweets related to traffic accidents are thus in small quantity. What’s more, most of them are confined to a small area and limited to a relatively short time interval and some researchers call them small-scale events (Schulz et al., 2013). Thus, the effectiveness and limitations of tweets
in detecting small-scaled events, especially the features of timeliness, accuracy, etc., should be explored and discussed. Second, another challenge in tweets lies in its inherent complexity and unstructured nature of data: language ambiguity (Chen et al., 2014b). The common methods in detecting the traffic-related events include support vector machine (D’Andrea et al., 2015; Schulz et al., 2013), natural language processing (Li et al., 2012; Wanichayapong et al., 2011), etc. which explore the semantic features in the keywords. However, as the context of a tweet is limited to 140 words and the tweet contents try to be concise, keyword detection is sometimes not sufficient for accurate automatic language processing. For example, “internet traffic is slow” and “internet shows traffic is slow” may deliver totally different information. To address above challenges, this dissertation explores and implements the association rules in the tweet contents for the traffic accident detection. Third, also due to the word limitation, some tweet contents that do not give enough descriptions to the incident types, even if some of the incidents may come from their suppositions. As indicated in Gu et al. (2016), the types of traffic incidents include car crash, construction work, bad weather, traffic congestion, etc. Further, all these types of incidents could incur the same congestion. The tweet users may only tweet about the traffic jam on the road.

2.3.2 Traffic surge and social event detection based on social media

Road traffic surge aggravates the jammed condition and worsens the level of service of road links. The consequences of traffic surge may vary, including traffic delay, fuel wasting, drivers’ frustration, etc. Some of the traffic surges may be accounted by the recurrent features of traffic patterns such as time-of-day characteristics and weekday-weekend differences. This kind of traffic increase is predictable in most cases, and people living by
usually get accustomed to it. Other kinds of traffic surge, which are more unpredictable and hazardous, may correlate with non-recurrent traffic patterns such as road accident, bad weather, malfunction of traffic signals, festival parades, etc. Figure 2-2 illustrates the differences between recurrent and non-recurrent traffic surge. The non-recurrent traffic patterns and corresponding traffic surge problems, which are caused by major social activities, are the main interests of this dissertation.

![Figure 2-2 Recurrent traffic fluctuations of hourly traffic occupancy in one sample detector](image)

For decades, the traffic surge problem, which potentially causes and even worsens traffic congestion, has been given much attention. The state-of-art studies attempt to unveil the correlation between traffic increase and other variables, and several explanations have been put forward, which are described as follows. Bando et al. (1995) found that there exists the congestion that is induced by a small perturbation without any specific origin such as a traffic accident or a traffic signal. Arnott et al. (2006) argued that cars cruising
for parking add to traffic congestion. Duranton et al. (2011) showed no evidence that the provision of public transportation affects vehicle kilometers traveled (VKT). Further, they proved that increased provision of roads or public transit is unlikely to relieve congestion. Anderson (2013) concludes that the cessation of transit service may increase average highway delay by up to 47%. Other studies even showed that higher congestion through restraining capacity for additional travel appears to be associated with the decrease in regional employment growth rates (Sweet, 2014). State-of-art studies investigate under which conditions and activities the traffic operations are influenced, and the traffic congestion deteriorates. However, the same as many other traffic problems that are closely related to human activities, the answers to correlation studies of the traffic congestion problems may be quite diverse.

Created in March 2006, Twitter founds a perfect stage of “We Media” and this makes possible the wide-range information retrieval about public activities from the massive majority of people. Over the past decades, social media has been validated useful to broadcast major events such as natural disasters (2013; Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events (Shirky, 2011), etc. The problems are whether it can detect the traffic surge and whether there exists any possible correlation between the tweets and traffic surge. Usually, the Twitter contents that are related to specific events will occur disproportionally frequently over certain time and space, and it is possible to make a direct connection between tweets and events like ceremony opening (Balduini and Della Valle, 2012), celebrity death or festival parades (Schwarz, 2012). However, the non-recurrent traffic surge may be quite different from that of other events. This is because a variety of public activities possibly incurs more on-road traffic and cause non-recurrent traffic
congestions. For example, on June 9th, 2013, an anomalous 10-mile traffic jam was detected on a major Southern California freeway. At the same time, the keywords “Obama” and “Impeach” occur much more frequently in the current tweets than the former ones (Giridhar et al., 2014). The contents that most people discuss and post on Twitter may imply a major social trend and bring more on-road traffic over specific time and space. In our study, we define **Twitter concentrations** as the tweets that involve a variety of traffic-related activities whose contents are widely created, consumed, distributed or shared. The goal of this dissertation is to explore the correlation between Twitter concentrations and traffic surge.

There are mainly three challenges to be addressed: The first challenge lies in how to quantify the traffic surge that may result from major public events. The time-of-day traffic data collected by loop detectors inheritably contain detection errors, and the traffic occupancy may fluctuate over time-of-day. Thus, it may not be easy to reasonably interpret the traffic data from the observations of one or two days and also not easy to determine the traffic surge according to data records previously collected. Also, traditional studies focused mainly on several intersections (Teodorovic et al., 2001) or corridors (Lan et al., 2008; Schoenhof and Helbing, 2007). In comparison, archiving, interpreting and summarizing the high-resolution traffic occupancy data in a large road network is quite challenging, considering the huge data size collected by large-scale fixed detectors. In our study, a clustering method is employed, and a detector-based probabilistic model to detect traffic surge is proposed.

The second challenge lies in the inheritable complexity and unstructured nature of tweet data: language ambiguity (Chen et al., 2014b) and how to exact the traffic-related
Twitter concentrations from the large collections of tweets is worth studying. We extract the Twitter concentrations by one or more keywords that make the tweets discriminatively different from that of others. The extracted tweets are further classified to label whether they are traffic-related or not. The prevailing methods of classifying tweets can be categorized into supervised and unsupervised techniques including Naïve Bayes classifier (Sankaranarayanan et al., 2009), online clustering (Phuvipadawat and Murata, 2010), support vector machine (Sakaki et al., 2010), hierarchical divisive clustering (Long et al., 2011), discrete wavelet analysis (Sakaki et al., 2010), continuous wavelet analysis (Cordeiro, 2012), decision trees (Popescu and Pennacchiotti, 2010) etc. The performance of these methods is partially decided by the applications and data sources. In this dissertation, we employed widely-accepted unsupervised and supervised learning techniques to classify the traffic-related tweets.

The third challenge is that public events that are reflected on the Twitter concentrations may exert different levels of influences on its surrounding traffic. Some of the activities may pose influence on more than one detectors, and some may be in effect within more than one hour. To properly interpret the traffic surge by the data collected from different locations and time periods, we can aggregate the data by two different statistics measures: mean and 75th percentile. These values can be compared with the traffic-related Twitter concentrations to explore their correlations with traffic surge.

2.4 Trip purpose inference

In the attempts of inferring the trip purposes, several strategies have been developed to alleviate the problems of decreasing responses including: increasing the number of
reminders, data fusion, more data-collection methods, etc. (Moiseeva et al., 2010). With the thriving growth of geocoded land use data and increasing popularities of mobile devices using satellite-based positioning system, a shift happens in the related studies from active individual or household travel survey to a more passive integral approach. For instance, the mobile phone records obtained from the mobile towers can reveal the human mobility pattern and trajectory features (Gonzalez et al., 2008; Zhang et al., An Empirical Study of Trip Purpose Inference with Connected Vehicles Trajectories, Land Use Data, and Social Media Data), inter-urban trip patterns (Kang et al., 2012), etc. Besides, the Global Positioning System (GPS) is also a well-accepted data source in extracting the information of trips or even tours: Wendy et al. (2009) combined the GIS data, GPS logs and the individual characteristics to interpret and validate the travel patterns; Anastasia et al. (2010) designed TraceAnnotator system that processes multiday GPS traces semi-automatically to impute transportation modes, activity episodes, and other facets of activity; An experiment in (2013) using the trajectories data recorded by a passive GPS summarize trip activities as “home”, “work”, “education”, “shopping” and “other”. Other GPS-enabled devices such as smartphones, on-board navigators can also record data effectively which is further employed to represent distinct mobility patterns across market segments typically relevant to transportation planning applications (Alexander et al., 2015). In most of the studies, the land use data are usually incorporated to assist the detection of trip purpose (Griffin and Huang, 2005; McGowen and McNally, 2007). For example, in (Alexander et al., 2015), trip origins and destinations are aggregated to a certain geographic scale to find the traces of the trip purposes from the land use classification; and the size of the areas used to aggregate trips is a very important factor.
For the trip information inference studies, there is still room to improve on the issues of accuracy, resolution, etc. for a more informative result. First, distributing GPS devices to the travelers is still an experimental approach and devices act as external rather than everyday necessary: the devices may be less taken care of in the long run making the data less accurate or even incomplete just like when using travel diaries. The GPS-based smartphone that passively collects location-time data is a better solution than the external GPS devices in the travel-survey studies (Stenneth et al., 2011). However, when come to combine the GPS records and the land use data, researchers may find land use category are too general to provide detailed useful trip information. Figure 2-3 shows a map including a trip end, the parcel of land use, and all commercial places within it. The category of the
land use is “Non-Residential Mixed Use”. As compared, the places within it give more detailed information. Second, the lack of historical massive activity information is another shortcoming for trip purpose studies in building an accurate model. As shown in Figure 2-3, when there are more than one “places” around a trip end, one may not find the exact destination of the traveler. (The “places” in this dissertation include commercial or non-profit organization that are open to the public including shop, market, company, organization, etc.) One possible solution is to infer the trip purpose through datasets that reflect the social trends of both the traveler himself and other people. However, as is said before, acquiring the detailed trip data from the crowds is always difficult and deficient.

This dissertation leverages geo-tagged Twitter data to infer trip purpose and fills the research gaps between reliable trip information and passively collected big data. With the increasing awareness of social networks and the rapid popularization of the smartphones, Twitter, the microblogging service, has received increasing attentions in recent years. Twitter establishes an online environment where the content is created, consumed, promoted, distributed, discovered or shared for purposes that are primarily related to communities and social activities, rather than functional task-oriented objectives (Gal-Tzur et al., 2014). Thus, not only time-location data but also the trip or location information are recorded through user-generated tweets. The massive collections of tweet data can generate more direct, representative and accurate information than that of the survey data or land use data as it collects the trends of location preferences and social awareness from the broad masses of the people. The data utility provided by Twitter is also promising: the corresponding time and location information along with geo-tagged Twitter can be of great research value and the resolution to pinpoint the locations can be as low as a few meters.
Previous social media studies (Lin et al., 2015; Ni et al., 2014, 2015; Zhang et al., 2016d; Zhang et al., 2016e) in transportation area mainly fall into two applications: traffic incident detection (Mai and Hranac, 2013; Wanichayapong et al., 2011; Zhang and He, 2016), and traffic prediction (Chen et al., 2016; Daly et al., 2013; Gal-Tzur et al., 2014). In other study areas, Twitter has been proved to be valid in the detection and broadcasting of social events including earthquakes (Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events (Shirky, 2011), etc. However, seldom do studies relate the social media to the trip purpose information inference.

There are major challenges to be addressed before the use of tweets in extracting useful trip information. First, the tweet data is mainly comprised of the inherently complex and unstructured word texts, and the language ambiguity (Chen et al., 2014b) in the tweet contents make them difficult to interpret. What’s more, as the context of a tweet is limited to 140 words and tweet users usually intend to be concise, the common methods in the language processing studies such as support vector machine (D’Andrea et al., 2015; Schulz et al., 2013), natural language processing (Li et al., 2012; Wanichayapong et al., 2011) may not be adaptive. Second, it is not an easy task to extract and further categorize the trip ends because not only the topics of posted tweet contents are mostly complicated and diverse but also that the information of activity location are embedded in the names that are not straightforward to be interpreted. For example, from tweets we found the place name “paint and pour” and one can’t identify it as the “private school” until searching it online. Third, same as other passively collected data, social media data generally lack ground truth of individual’s travel modes and trip purpose information. Inferring the ground truth of a personal trip is very difficult due to personal privacy. However, this dissertation promotes
the study by verifying that the trip end GPS locations inferred from social media can match that of the actual facilities and thus can be taken as trip purpose information.
CHAPTER 3 SPATIAL-TEMPORAL TRAFFIC FLOW PATTERN IDENTIFICATION AND ANOMALY DETECTION

In this chapter, we unveil the important features of traffic flow patterns separately from spatial and temporal perspectives. The details are elaborated in the regional traffic pattern identification paper (Zhang et al., 2016c) published in 2016. From the spatial perspective, the problem is how to measure and evaluate the anomaly degree of a road link or intersection approach as compared to others. This relates to the identification of recurrent traffic flow patterns and provides insights in how to constantly identify problematic locations within a certain scale of the road network. From the temporal perspective, the problem is how to measure and evaluate the abnormal time-of-day traffic flow patterns as compared to historical records. This relates to the non-recurrent traffic flow patterns and helps traffic operators learn why traffic goes to the extreme in one day whereas those in other days are relatively normal. What’s more, both spatial and temporal patterns can be interpreted in different geographic and time scales.

3.1 Data preprocessing

The first task is to discretize the traffic-related features and for each feature and put the traffic data describing the similar traffic conditions into the same category based on its physical meanings. The features we need include traffic flow, traffic occupancy or other geographic information and categorization process requires the pre-definition of the bounds of the features.

For geographic information, we studied the county, sub-region and intersection level all of which are discretized information. The county that a detector belongs to has three
categories: Loudoun County, Prince William County, Fairfax County. The sub-region information is further manually divided into more than 21 sub-regions by the Virginia Department of Transportation (VDOT) for traffic operation purposes. As to the intersection level, there are two categories: “Major” and “Non-Major”. The “Major” intersections are those whose major roads are arterials, and the “Non-Major” intersections are those whose major roads are collectors or local roads. The levels of geographic scales can be visualized in Figure 3-1.

![Figure 3-1](image)

Figure 3-1 Three geographic scales to identify both the spatial and temporal traffic flow pattern within a road network

For the traffic information, the traffic flow per lane is categorized into three levels: Level A: [0, 700), Level B: [700, 1200), Level C: [1200, 2000] according to the service flow rate (AASHTO, 2001; FHWA, 2000). The unit is vehicles per hour (vph). As the peak-hour link flow on a typical arterial may not exceed a certain limit, we exclude these
data with flow higher than 2000 vph. It is worth mentioning that some studies classified
the traffic flow into four stages: blocking flow, crowded flow, steady flow and unhindered
flow (Rong et al., 2013). This is almost the same when the first two groups are merged into
one. Given that the occupancy is linearly related to the density, the categorization should
not only refer to (AASHTO, 2001; FHWA, 2000) but also the study of the flow-density
relationship in the past few decades. All of the previous studies accordantly indicate that
same volume of traffic flow may correspond to twofold different density or occupancy
values. Then, according to the definition of the level of service (LOS) from A to E, the
traffic occupancy is categorized into five different levels: Level 1: [0, 1/17), Level 2: [1/17,
1/9), Level 3: [1/9, 1/7). After occupancy reaches 1/7, the traffic flow capacity is almost
reached, and we introduce another 2 categories to distinguish 2 degrees of traffic jam: Level
4: [1/7, 1/2), Level 5: [1/2, 1]. After conversion, both the traffic flow and occupancy that
are collected in the 15-minute interval have been discretized. The discretization is the initial
and also the crucial part of regional traffic flow pattern identification. The boundaries of
each level in a category should follow the directions of level of service (LOS) (AASHTO,
2001; FHWA, 2000) because it is widely accepted that traffic data falling into the same
LOS follow the same traffic flow pattern. Level setting is crucial: if we set fewer levels to
a category, the traffic flow patterns can not be properly distinguished. If we set more levels
to the category, some of the free-flow traffic data may be identified as anomalies by the
compression method because sometimes near-zero flows could be rare in a congested road
network.

The geographic and traffic information constitute a database as the input of our
algorithm. Table 3-1 briefly lists all the traffic-related features in this chapter.
Table 3-1 Feature table employed in the regional pattern identification study

<table>
<thead>
<tr>
<th>Features</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>Loudoun County; Prince William County; Fairfax County;</td>
</tr>
<tr>
<td>Sub-region</td>
<td>Divided into 22 as shown in Figure 3-1</td>
</tr>
<tr>
<td>Intersection level</td>
<td>Major; Non-major;</td>
</tr>
<tr>
<td>Traffic flow</td>
<td>[0, 700); [700, 1200); [1200, 2000);</td>
</tr>
<tr>
<td>Traffic occupancy</td>
<td>[0, 1/17); [1/17, 1/9); [1/9, 1/7); [1/7, 1/2); [1/2, 1];</td>
</tr>
</tbody>
</table>

For the time information, we do not include any of them as features of the database. Instead, we try to aggregate the database according to the different time scales to identify the pattern differences in different time periods. Furthermore, we also consider the patterns on weekdays and weekdays separately in section 4.

3.2 Method

3.2.1 Dictionary-based compression

After categorization, a raw database ending up with a rather “redundant” set of patterns has been created. The word “redundant” refers to that the database can be further compressed and generate useful knowledge. To exemplify the dictionary-based compression technique, we continue with a simple illustration. We assume to have a database $D$ with 3 different features: Flow ($F$), Occupancy ($O$), Intersection ($I$). In each feature, there is a domain of possible values: $dom(F)=\{F_1\ldots F_3\}$, $dom(O)=\{O_1\ldots O_5\}$, $dom(I)=\{I_1\ldots I_2\}$. The combination of all features in each 15-minute time period is taken as a database (DB) pattern: $DF_i=\{F_j, O_k, I_l\}$. Theoretically, the domain of $DF_i$: $dom(DF_i)$ is all the possible combinations of features of the database.
The next step is to build a suitable pattern table (PT) to compress and encode the features of the database. Table 3-2 shows an illustrative example of the Database Table and its PT. There are two columns in a PT. The first column is the PT pattern column. The domain of PT pattern: $dom(PF)$ can be different from that of DB pattern $dom(DF)$. The PT patterns can be the combination of any feature values but all PT patterns are included in $dom(DF)$: $\forall PF \in dom(DF)$. The second column is the number of occurrences of each PT pattern in the database. The PT then performs as a code dictionary and the process of converting the DB patterns into combination of PT patterns is called dictionary-based compression. One can see that after conversion, the previous larger database is compressed and encoded into a smaller one and the usage of PT patterns differs from each other in the PT Table. One can assume that different pattern tables result in different PT patterns and thus different usages.

Table 3-2 An illustrative example of database table and pattern table

<table>
<thead>
<tr>
<th>Database table</th>
<th>Pattern table</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB pattern $(DF_i)$</td>
<td>PT pattern included</td>
</tr>
<tr>
<td>F</td>
<td>O</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level B</td>
<td>Level 4</td>
</tr>
<tr>
<td>Level B</td>
<td>Level 3</td>
</tr>
</tbody>
</table>
For DB pattern encoding, given the usage, one can compute the optimal lengths of the code words to encode the patterns according to optimal prefix code (Rissanen, 1978). The length of PT pattern \( PF_i \) in a certain pattern table is defined as:

\[
L(usage (PF)|PT) = - \log \left( \frac{usage (PF)}{\sum_{PF_i \in PT} usage (PF_i)} \right)
\]  

(3-1)

It is worth mentioning that the base of all logarithms in this chapter is 2 (Agrawal and Srikant, 1994). Then, the length of DB pattern \( DF_i \) is calculated as the sum of the lengths of all PT patterns it contains. We find the best set of \( PF \) that can make up \( DF \). In one \( DF \), any component \( PF \) cannot cover other \( PF \)s.

\[
L(DF|PT) = \sum_{PF \in DF} L(usage (PF)|PT)
\]  

(3-2)

The length of DB is the sum of the lengths of all DB patterns the database contains.

\[
L(DB|PT) = \sum_{DF \in dom(DF)} L(DF|PT)
\]  

(3-3)

For encoding the pattern table, we still need optimal prefix code (Akoglu et al., 2012). There are two parts involved in the length of PT. The first part is the sum of lengths of all PT patterns; the second part is the sum of lengths of all singleton items in each category in DB. Define \( l \) as all the singleton items in DB, \( c \) as the total count of singleton items and \( r_i \) is the count of the \( i \)th singleton item. For example, in Table 3-2, \( c \) is equal to 24, \( r_i \) of the singleton item “Level A” is 6.

The length of the PT table is defined as:
The length of a DB pattern code plays a significant role in anomaly detection. To save storage space, patterns that occur more frequently result in smaller \( L(PT) \), which are regarded as the normal ones, whereas those occurring less frequently are taken as the abnormal ones. For example, in transportation study, the unexpected non-recurrent traffic congestions are treated as anomalies and the severity of the anomalies can be quantified by the length of the DB patterns.

### 3.2.2 Minimum Description Length principle

One can see from Section 3.2 that given a database of observations, the lengths of both DB and PT are totally decided by the selection of PT. One principle, called Minimum Description Length (MDL), should be followed to select a suitable PT to compress and encode the database. The MDL principle identifies the best PT which minimizes the description length:

\[
\mathcal{L} = L(DB|PT) + L(PT) \tag{3-5}
\]

The MDL principle requires us to find the description length of the shortest codes for the actually observed sequence (data), rather than a mean length (Barron et al., 1998). Also, one can see that a complex PT with diverse PT patterns can compress the DB very well and thus leads to a small \( L(DB|PT) \). The side effect is that it will also result in a large \( L(PT) \). MDL principle attempts to balance the complexity of PT and its fit to DB. Therefore, the PT that can provide the shortest description length of Eq. 5 is the best pattern table that compresses the database.

To find the best PT Table, we employ a heuristic search algorithm as follows.
**Algorithm:** Dictionary-based compression

**Input:** Database with $n$ rows and $m$ categories

**Output:** A PT table and the length of each pattern

**Build** the initial PT table and all PT patterns $PF_i$ are singleton items of features in DB

**Compute** the initial description length $L_0$, the optimal length $L = L_0$

**Implement** the Apriori algorithm to find all frequent items $FI$ whose frequency is higher than a threshold $T$, these frequent items constitute a set $S$

**Repeat**

- for $FI_i$ in $S$
  - Put $FI_i$ into the PT table
  - Compute the current description length $L_i$
  - If $L_i < L$
    - $L = L_i$
    - remove $FI_i$ from $S$
    - add $FI_i$ into PT table
  - else
    - remove $FI_i$ from $S$

until $|S| = 0$

The proposed search algorithm can be interpreted in the following steps:

**Step 1:** All possible singleton items in DB are taken as the PT patterns in the PT table.

For example in Table 3-2, the PT patterns are (“Level A”, “Level B”, “Level 2”, ”Level 3”, “Level 4”, “Major”, “Non-major”). According to these PT patterns, one can calculate the initial description length and take it as the current length.

**Step 2:** We implement the Apriori algorithm (Agrawal and Srikant, 1994) to find all frequent items that are the combinations of one or more singleton items. We order the items according to their frequency and choose those whose frequency is higher than a threshold. These “frequent” items are the potential candidates for PT patterns. For example, in Table
3-2, the combination of “Level A” and “Level 2” has the highest frequency of 6 and they are put on the list of candidates.

Step 3: We add the most frequent combination from the list of candidates into the PT table and recalculate the description length. If the recalculated length is smaller than the current one, we add the new item to the PT table and change the length of each pattern. If the recalculated length is larger, we keep the previous PT table and score. For example, in Table 3-1, after calculation, it is found that adding the item “Level A/Level 2” to the PT table can reduce the description length. Therefore, the item “Level A/Level 2” is chosen and added to the PT table.

Step 4: We remove the item in Step 3 from the candidate list and continue with the next candidate until there is no candidate left on the list.

It is worth mentioning that the threshold $T$ for the candidate item is chosen as 30% of the total count of DB patterns in this chapter. One can still lower the threshold but this may lead to more computation and relatively less improvement.

### 3.2.3 Anomaly degree

In Section 3.2.2, the MDL principle finds the best pattern table, and we proceed to derive a normalized anomaly degree that can characterize the traffic patterns.

Given a PT, the length of a traffic pattern $PF$ is:

$$L(DF|PT) = \sum_{PF \in DF} L(usage (PF)|PT)$$

(3-6)

$L(DF|PT)$ indicates the anomaly degree of a traffic pattern. The higher the length is, the closer the traffic pattern length is to the upper bound, and consequently the more abnormal the traffic pattern should be. Theoretically, when the DB table has only one kind
of $DF$, there should also exist only one kind of $PF$ equal to $DF$ in $PT$. The lower bound of $L(DF|PT)$ is inevitably 0. However, the upper bound should be bounded by a certain value.

We continue to approximate the upper bound by estimating the possible largest values of the pattern length.

**Definition 1:** Given the number of DB patterns, the approximate upper bound of $L(DF|PT)$ is defined as:

$$U = \tau \cdot \sum_{i \in \text{dom}(i)} -\log \left( \frac{1}{N} \right)$$

(3-7)

Where $N$ is the total number of rows in DB, $\tau$ is a discount factor.

**Remark 1:**

$$L(DF|PT) = \sum_{PF \in DF} L(usage (PF)|PT)$$

$$= \sum_{PF \in DF} \max \{L(usage (PF)|PT)\}$$

$$= \sum_{PF \in DF} \max \left\{ -\log \left( \frac{usage (PF)}{\sum_{PF \in PT} usage (PF_i)} \right) \right\}$$

Two cases need to be considered separately.

Case 1: There exists only singleton items in PT table, then the highest values of $-\log \left( \frac{usage (PF)}{\sum_{PF \in PT} usage (PF_i)} \right)$ for each PT feature in DF should be $usage (PF) = 1$ and $\sum_{PF \in PT} usage (PF_i) = N$ that is $-\log \left( \frac{1}{N} \right)$.

Case 2: There exist PT patterns that are combinations of 2 or more singleton items. For PT patterns with more than one singleton items, $-\log \left( \frac{usage (PF)}{\sum_{PF \in PT} usage (PF_i)} \right) = \sum_{i \in \text{dom}(i)} -\log \left( \frac{1}{N - \sum_{i \in \text{dom}(i)} M_i} \right)$ where $M_i$ equals to the number of combined PT patterns.
that contain the singleton items in $i$th feature. Also, $\sum_{i \in dom(i)} - \log \left( \frac{1}{N - \sum_{i \in dom(i)} M_i} \right) \leq \sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right)$.

However, after our examination, the circumstance that all features with usage ($PF$) = 1 is really rare and $\sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right)$ is too large for most of DB table. To make the anomaly degrees from different DB comparable and also make it less conservative, we introduce $\tau$ as a discount factor. Therefore:

$$L(DF|PT) \leq \tau \cdot \sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right)$$

In this chapter, $\tau$ is set as 0.85. Defining $U$ allows us to normalize the anomaly index to be a fraction number. We further derive the anomaly degree to evaluate the performance of traffic patterns:

$$D(DF_i|PT) = \frac{L(DF_i|PT)}{U}$$

(3-8)

Where $D(DF_i|PT)$ is the normalized anomaly degree of the $i$th traffic pattern in the database. Note that since $U$ is an approximated value, $D(DF_i|PT)$ could possibly exceed 1.0.

One can see that the anomaly index is a comparable index and can identify the least frequent traffic flow patterns as the abnormal ones. This index is valid because given the vast study area in NOVA, the places that experience abnormal traffic operations such as traffic jam, traffic accident, festival parades etc. take up only a small portion of the whole area. This rule also applies to the abnormal time periods of a given location. Furthermore, from Eq. 5, this index includes more than one traffic-related metrics including the geographic information. Through compression, this inclusion of geographic information can decrease the index where the traffic is always jammed (abnormal) as well as increase
those where the traffic is unexpectedly abnormal. It can be concluded that the anomaly index can compare the overall performance of traffic operations in different places or different time periods. The lower the value is, the less abnormal the traffic condition and the better traffic performance should be. The absolute value also poses significance on the traffic performance evaluation. Our empirical examination shows that most of the values are below 0.7. The reasons anomaly index higher than 0.7 may be another interesting topic and not be covered in this chapter. What interests us is the performance measures in different spatial and temporal perspectives.

Here we define two anomaly index: spatial anomaly degree and temporal anomaly degree:

- For spatial anomaly degree, the DB table is the set of traffic flow patterns in the whole region over the same time period.

- For temporal anomaly degree, the DB table is the set of traffic flow patterns in one location over the same time period on different days.

One can see that both spatial and temporal anomaly degrees aim at capturing the traffic flow patterns that are different from the rest majority. The spatial anomaly degree may indicate locations that have much heavier traffic than their surrounding ones such as overcrowded traffic in afternoon peak, typically caused by recurrent commuter traffic. The temporal anomaly degree may indicate the time periods in one day when traffic is usually lighter or heavier than other days; This index is important to identify non-recurrent traffic flow pattern that may be caused by unexpected incidents vehicle breakdowns, crashes, the closure of road links, festival parades, malfunction of the traffic signals, etc. Previous
studies show that the non-recurrent traffic flow pattern is even more important because it accounts for more than 50% of the total motorist delay (Chin et al., 2002).

The causes of traffic accidents are various and the corresponding strategies should be devised accordingly which may include traffic signal control (Asamoah, 2014; Ding et al., 2015; He et al., 2011, 2014; He et al., 2016), urban design (Saelens et al., 2003), work zone control (Beacher et al., 2004), incident management (He et al., 2013; Ozbay and Kachroo, 1999), etc. These will not be covered in this section. Instead, we focus on the anomaly detection method and its applications.

3.3 Numerical examples

In this section, we implement the proposed traffic flow pattern identification algorithm in NOVA area. We will first explore the physical meaning of the anomaly degrees by comparing results between our compression-based method and other pattern recognition methods, then reveal the spatial and temporal traffic flow pattern characteristics and finally prove the validity of our method by a case study.

3.3.1 Spatial regional traffic flow pattern

The spatial regional traffic flow pattern helps identify the abnormal locations that display different characteristics from the rest locations. It is a straight-forward tool for traffic operators to compare the traffic flow patterns in different locations over the same time period. We first calculate the anomaly degrees of the traffic flow patterns in the detectors and further derive the anomaly degrees of the intersection by taking the 90th percentile of anomaly degrees of all detectors that belong to the intersection. This is reasonable because the overall performance of the intersections can be worsened by one or
two jammed directional links. Figure 3-2 shows the heat map of spatial regional traffic flow patterns at intersection level in PM peak (5:00-6:00 p.m.) separately on a weekday (Feb 20, 2014) and a weekend (Mar 22, 2014). The scale of the color in different plots is set to be equal in order to make the anomaly degrees comparable. It is concluded that the red (dark) dots indicate a totally different traffic flow pattern from the rest majority and the redder the dots are, the more abnormal the places should be. The detector anomaly degrees of two most abnormal intersections are displayed below the heat map.

Here we have several interesting observations. First, during PM peak, the overall anomalies mainly concentrate in a few intersections, indicating the traffic jam and also the lack of road facilities. One can see some distinct red corridors, such as Leesburg Pikes, which goes from northwest to southeast. This indicates that certain roads in the network experience more severe conditions than that of others. In addition, one can say that certain regions are more likely to be jammed together (Banai-Kashani et al., 2011). The numerical results of some studies revealed that different traffic flow patterns could emerge depending on the (different) observed time-scale (Lan et al., 2008). In this chapter, however, there is almost no difference in geographically distributed features of anomaly locations on weekdays or weekends during PM peak.

To identify the day-to-day fluctuations of spatial traffic flow patterns, we randomly selected two intersections in our studied region, which are located in different counties. The time period is also set in PM peak (5:00-6:00 p.m.). From Figure 3-3, most of the traffic flow patterns do not show large fluctuations from January to September. This means that the traffic performance of spatial traffic flow patterns for one intersection over certain time periods in a network remain almost constant in different days.
Figure 3-2 spatial regional traffic flow pattern identification during PM peak (5:00-6:00 p.m.) on (a) Feb 20, 2014 (Weekday) and (b) Mar 22, 2014 (Weekend). The anomaly degrees of all detectors (x-axis of the bar plots) in two most abnormal intersections and their locations are shown in the bottom of the plot.

The unchangeable nature of the traffic flow patterns may not be true for those traffic flow patterns in one day during different time periods. Due to the fact that we have more
than 1200 signalized intersections, aggregated anomaly degrees in different sub-region may be more helpful in identifying the pattern shift. Table 3-3 presents the anomaly degree results of three sub-regions. The anomaly degrees increase as the geographic locations gradually move closer to the District of Columbia. Also, from Figure 3-4, one can see a clear time-of-day pattern shift in almost all sub-regions, a trend of regional anomaly degrees decreasing from AM peak to noon and increasing back to PM peak. In addition, the ranks of the sub-regions remain almost the same.

The spatial traffic flow pattern identifications and the corresponding anomaly detection may provide useful information for traffic operators. The places with higher anomaly degrees may experience heavy traffic and elongating travel time. As some studies on the congested traffic states that typical kind of traffic congestion depends on the specific freeway (Treiber et al., 2000), the abnormal traffic flow patterns also emerge on specific road lanes. The causations may be the layout of the road network or other infrastructure reasons and it should be an interesting topic in further studies.
Figure 3-3 Anomaly degrees of spatial traffic flow patterns of two intersections on PM peak (5:00-6:00 p.m.) in different days. The first intersection (a) is Richmond Hwy. and Sherwood Hall Ln.; the second intersection (b) is Lee Hwy. and ramp of W Ox Rd.

Table 3-3 Anomaly degrees of spatial traffic flow patterns in 3 sub-regions in different time periods on weekday February 2nd, 2014.

<table>
<thead>
<tr>
<th>Region ID</th>
<th>AM peak (7:00-8:00)</th>
<th>Noon (11:00-12:00)</th>
<th>PM peak (17:00-18:00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.725</td>
<td>0.668</td>
<td>0.722</td>
</tr>
<tr>
<td>19</td>
<td>0.686</td>
<td>0.641</td>
<td>0.673</td>
</tr>
<tr>
<td>16</td>
<td>0.593</td>
<td>0.561</td>
<td>0.590</td>
</tr>
</tbody>
</table>
3.3.2 Temporal regional traffic flow pattern

The temporal regional traffic flow pattern aims to identify the abnormal patterns over certain time periods in one day that display different pattern characteristics from the other days. Same as the spatial traffic pattern identification, we first calculate the anomaly degrees of the detectors and aggregate them into the anomaly degree of the intersection by taking the 90th percentile. We also consider weekdays and weekends separately.

Figure 3-5 shows a quite different heat map of temporal traffic flow pattern than that of spatial traffic patterns. As one can see, there is no clear corridor pattern of abnormal intersections. The occurrences of abnormal traffic flow patterns are quite random. If the criterion of anomaly degree is set as 0.8, less than 10 percent of the intersections are abnormal, whereas more than 90% of the intersections operate pretty much the same as most of the other days. For two most abnormal intersections, the anomaly degrees of the detectors in the intersection differ greatly from each other, and the traffic flow pattern of an intersection may deteriorate by one or two abnormal detectors.
We further select two intersections and examine the temporal traffic flow pattern fluctuations over the PM period in different days. From Figure 3-6, we can see that the abnormal degree oscillates in different days. The rare peak points are identified as abnormal patterns, whereas most of the rest represent normal patterns. The temporal abnormal pattern could be observed during both weekends and weekdays and these can help the traffic operators identify the unexpected events in certain locations.
Figure 3-5 Temporal regional traffic flow pattern identification during PM peak (5:00-6:00 p.m.) on (a) Feb 20, 2014 (Weekday) and (b) Mar 22, 2014 (Weekend). The anomaly degrees of all detectors (x-axis of the bar plots) in two most abnormal intersections and their locations are shown in the bottom of the plot.

Figure 3-6 Anomaly degrees of temporal traffic flow patterns of two intersections on PM peak (5:00-6:00 p.m.) in different days. The first intersection (a) is Seminary Rd. and South George Mason Rd.; the second intersection (b) is Rixlew Ln. and Wellington.

Table 3-4 shows the anomaly degrees of temporal traffic flow patterns in different sub-regions in different time periods on the same weekday. Unlike the spatial flow pattern, the temporal anomaly degrees in the sub-regions do not show a clear trend in different time
periods as shown in Figure 3-7. This observation also proves that the temporal flow pattern has little to do with geographic locations.

One can see that the identification of temporal traffic flow pattern may find the unexpected anomalies that appear quite randomly in both time and space. It can be assumed that under this circumstance, the traffic operations or travel time over certain links may be greatly influenced.

Table 3-4 Anomaly degrees of temporal traffic flow patterns in 9 sub-regions in different time periods on the same weekday

<table>
<thead>
<tr>
<th>RegionID</th>
<th>AM peak (7:00-8:00)</th>
<th>Noon (11:00-12:00)</th>
<th>PM peak (17:00-18:00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.220</td>
<td>0.178</td>
<td>0.187</td>
</tr>
<tr>
<td>19</td>
<td>0.225</td>
<td>0.175</td>
<td>0.227</td>
</tr>
<tr>
<td>16</td>
<td>0.331</td>
<td>0.205</td>
<td>0.212</td>
</tr>
</tbody>
</table>

Figure 3-7 Temporal anomaly degrees in three different time periods for different sub-regions (each link represents a sub-region).
3.4 A case study

Serving different purposes, the spatial and temporal traffic flow patterns in one location may display different features. These features can jointly explain the changes of traffic flow patterns in two different perspectives. Figure 3-8 illustrates anomaly degrees of all detectors over three time periods: AM peak (7:00-8:00 a.m.), Noon (12:00-13:00 p.m.) and PM peak (16:00-17:00 p.m.), across two days, Feb 20, 2014 (weekday) and Mar 22, 2014 (weekend). One can see that anomalies may coexist both in time and space even though the situation is very rare. The correlation between the spatial and temporal anomaly degrees is not high. This means the spatial and temporal anomaly may not possibly influence each other and should be explained separately. The spatial patterns compare the traffic flow patterns across the geo-spans, whereas the temporal ones compare those across time-spans. Thus, if a location is congested all the time as compared to other locations in the network, it is abnormal in spatial traffic flow pattern; but it is possible that these such congested conditions are only normal day-to-day recurrent pattern if compared with the historical traffic flow patterns during the same time-of-day. Therefore, the spatial anomaly does not necessarily indicate temporal anomaly and vice versa. Combinations of both traffic flow pattern identifications can enhance our understanding of the regional traffic patterns.

We implement the proposed method to study the changes of traffic flow patterns of the new 11-mile extension of Silver Line, a subway line of Washington Metro. The extension consists of 5 exclusive new stations, which began service on July 26, 2014. It connects with the Orange Line at East Falls Church Station and extends to the west business district. 4 out of 5 stations are located in the densely populated area of Tysons Corner and may
bring impact to the local traffic and the regional traffic flow pattern. One can refer to Table 3-1 for feature and pattern definitions used in this case study.

![Spatial and temporal regional traffic flow pattern of all detectors on Feb 20, 2014 (Weekday) (a) 07:00-08:00 a.m., (b) 12:00-13:00 p.m. and (c) 17:00-18:00 p.m. and Mar 22, 2014 (Weekend) (d) 07:00-08:00 a.m., (e) 12:00-13:00 p.m. and (f) 17:00-18:00 p.m.](image)

Figure 3-8 Spatial and temporal regional traffic flow pattern of all detectors on Feb 20, 2014 (Weekday) (a) 07:00-08:00 a.m., (b) 12:00-13:00 p.m. and (c) 17:00-18:00 p.m. and Mar 22, 2014 (Weekend) (d) 07:00-08:00 a.m., (e) 12:00-13:00 p.m. and (f) 17:00-18:00 p.m.

We calculate the anomaly degrees of spatial and temporal traffic flow patterns of all detectors near the metro stations separately and then take an average of the anomaly degrees for each intersection. To make a clear comparison, we further calculate the ratios of the degrees after and before the Silver Line extension. Still, we consider the case on weekdays and weekends separately. Figure 3-9(a) shows the geographic information of the
sub-region containing these 4 metro stations. In the same geographic area, Figure 3-9(b)-(e) shows the ratios of spatial and temporal anomaly degrees before and after case separately on weekdays and weekends.

It will be helpful to examine the anomaly degrees together with the geographic-distributed features of the study areas. From Figure 3-9(a), the metro stations are evenly distributed along the road line of Route 7 and Chain Bridge Rd and one of the metro station (Greensboro metro station) is almost located at the intersections of two roads. On both arterials, one can see an overall increase of anomaly degrees both in spatial and temporal traffic flow patterns on weekdays and weekends after the Silver Line extension. It shows that new metro stations attract more commute traffic and entertaining traffic (for Tysons Corner Center in the red area). Besides these two arterials, other intersections, especially those collectors and local roads located within the commercial land, do not show an identical increase or decrease in different spatial-temporal perspectives. Some places even show a clear decrease of anomaly degrees which means less traffic jam is observed. As a conclusion, the newly built metro station may change the spatial and temporal traffic flow patterns. Potentially it worsens the traffic on major arterials and alleviates that of collectors and local roads.
Figure 3-9 Map of the sub-region; Ratios of spatial anomaly degrees between after and before Sliver Line extension in the same area (x-axis and y-axis are longitude and latitude,
respectively) on (b) weekdays and (c) weekends; Ratio of temporal anomaly degrees between after and before Sliver Line extension on (d) weekdays and (e) weekends.

For the spatial traffic flow pattern part, it should be noted that the metro stations are coupled with several newly-built transit lines designed to connect new Silver Line rail travel (WMATA, 2014). Most of the transit lines stop at the arterials instead of collectors or local roads. What is more, new parking lots are open together with the metro station and their locations are mostly near the arterials. The newly-built metro station may change both the land-use features and the trip mode and in turn, change the spatial traffic flow patterns. Over a certain time period, some links become more crowded and others become relatively less. One can say new transit systems possibly change the recurrent traffic flow patterns within a geographic scale. For the temporal traffic flow pattern part, these changes are even more apparent. The after-case pattern can be taken as the minority and detected just because the time period of after-case pattern is from July 26th to August 31st which is only a small portion of the total record period. Unlike spatial patterns, temporal traffic flow patterns capture the changes of recurrent traffic patterns as with the changes of transportation systems (e.g. new subway stations). It is worth mentioning that with longer observation time periods, the recurrent pattern changes will become stable and the new patterns will be regarded as normal ones. In sum, the spatial and temporal traffic identification reach conclusions in different ways and support decision making for transportation planning and management.
3.5 Comparisons

3.5.1 Comparisons between different discretization settings

We further conduct a study with the proposed methods by different discretization boundary settings of flow and occupancy in calculating the temporal anomaly degree.

We present the results in a flow-occupancy plot, coming after the concept of the fundamental diagram (Geroliminis and Daganzo, 2008), which assumes that flow and density (occupancy) follow a concave relationship. We randomly select one detector located on different links and conduct a temporal traffic flow pattern identification on these detectors. The DB is the set of traffic flow patterns (here only traffic flow and occupancy) over the same 15-minute period on different weekdays (weekend data are excluded).

![Flow-Occupancy Plot](image)

Figure 3-10 Examples of anomaly degrees in one detector during 17:00~17:15 p.m. on different weekdays. Obvious abnormal dots are circled. The boundary settings are defined as follows, (a) the same setting as introduced in Section 3; (b) different levels when discretizing traffic flow to [0, 500), [500, 100), [1000, 2000]; (c) different levels when discretizing traffic occupancy to [0, 20%), [20%, 40%), [40%, 60%), [60%, 80%), [80%,100%].
As shown in Figure 3-10 (a)-(c), the darker red dots indicate more abnormal traffic flow patterns. One can see from Figure 3-10 (a) that the traffic flow is around 500~700 during the peak hours. As compared, those traffic flow data lower than 500 with small occupancy are more abnormal. This is reasonable because relatively low traffic flow and occupancy happen during a rush hour which makes the traffic flow patterns on those days quite different from other days. For those isolated traffic data whose occupancy is apparently higher than 0.15, the patterns are also taken as the anomalies. The observations by the compression method in Figure 3-10 (a) coincide with our intuitive perceptions based on the fundamental diagrams. The results based on the even discretization in traffic flow and occupancy are shown in Figure 3-10 (b) and (c), respectively. In Figure 3-10 (b), with the upper bound for free flow is set to be 500, the anomaly degrees for those traffic flow patterns are relatively higher. In Figure 3-10 (c), the high upper bound (20%) for the first occupancy level already includes the majority of the traffic occupancy data. Manually changing the boundary settings may come up with totally different results and these results may not conform to the current LOS requirements.

Figure 3-11 (a)-(f) shows the effects of merging and dividing of original discretization levels on the anomaly degrees into two different detectors. The anomalies in two detectors can be explained from different flow and occupancy interpretation. In the first detector, the anomaly data are due to high occupancy while in the second one, those are mostly due to low traffic flow. Their results show that merging levels may decrease the anomaly degree values while dividing may do the reverse. This can be explained by the fact that a smaller number of levels corresponds to a smaller value of \( L(usage (PF)|PT) \) in Equation 6 and in turn a higher value of anomaly degree. Merging and dividing the levels may over or
underestimate the anomaly degrees but the ranks of each traffic flow pattern shall not change apparently. However, it can be assumed that if the size of the datasets is small, the effects may be quite unpredictable as smaller sample size may be less likely to uncover the regularities in the datasets.

Figure 3-11 Examples of anomaly degrees in two detectors during 17:00~17:15 p.m. on different weekdays with different boundary settings. The results of the first detector are shown in (a)-(c) and those of the second are shown in (d)-(f). (a) and (d) follow the boundary setting \([0, 1/17, 1/9, 1/7, 1/2, 1]\) as introduced in Section 3; (b) and (e) merge some occupancy categories into \([0, 1/17, 1/7, 1]\); (c) and (f) further divide the occupancy categories into \([0, 1/28, 1/17, 1/12, 1/9, 1/7, 1/2, 1]\).
### 3.5.2 Comparisons with another anomaly detection method

We further compare our proposed method with long-term anomaly detection method proposed by Vallis et al. (2014). The Vallis’ model is built on a daily basis and requires the detection of the trend component. It can detect anomalies according to the periodic features of the time-of-day traffic data. The method employs time series decomposition and robust statistics for detecting anomalies. It builds on the generalized Extreme Studentized Deviate test (ESD) (Rosner, 1983) which assumes data set comes from a normal-distributed population. According to Zhang et al. (2016e), the traffic occupancy over a certain period may follow a normal distribution. Thus, the method can potentially detect the traffic flow pattern anomalies with high precisions. Figure 3-12 shows the results on the same dataset used in Section 4. For display purpose, we only present the results of the time period of 17:15 ~ 17:30 p.m. on different days. Both methods present different anomaly detection results making it worthwhile discussing. In contrast, compression-based method emphasizes on the features over certain time periods on different days.

Figure 3-12 (b) and 13(c) present the data over the same peak period on different days. The anomalies detected by our method can be better explained. Two extremely low traffic flow and occupancy (shown as blue triangles) are taken as anomalies because they indicate traffic conditions that are substantially different from the majority of other peak periods. In contrast, Vallis’ model outputs three anomalies, which are shown as red rectangles. Two of three are with high flow and high occupancy, which are quite normal for peak periods. Therefore, temporal anomalies are not fully captured by Vallis’ model. Compared with Vallis’ model, the compression-based method finds these abnormal points partially or even wholly decided by the frequencies of their appearances. It opens a new window in detecting
the abnormal traffic flow pattern through the massive collection of traffic-related data. One can see that the compression technique identifies the abnormal points from a different perspective. The effectiveness of these methods can be further explored for more specific purposes of traffic flow pattern identification.

Figure 3-12 Temporal anomaly results by the proposed compression-based method, (b) traffic flow and (c) traffic occupancy over time period 17:45 ~ 18:00 p.m. in different weekdays. Red rectangles indicate the anomalies detected by (Vallis et al., 2014); blue triangles indicate those detected by compression-based method.
3.6 Conclusions and discussions

As a first attempt, this chapter leverages dictionary-based compression theory for regional traffic flow pattern identification and anomaly detection within a large-scale traffic network. Spatial and temporal anomaly degree indices are derived to describe and quantify the abnormal traffic flow pattern. Empirical results using traffic flow and occupancy show its adaptability in the transportation field. There are three major advantages to this approach that should be highlighted:

- Effective reduction of dimensionality: two or more traffic-related metrics are interpreted together which prevents the complex definition of several different variables and parameters. The features of traffic flow patterns can be quantified as a low-dimensional index.
- Clear representations of geographic-distributed features of spatial and temporal traffic flow patterns: In the heat map of our study area, anomaly locations can be easily identified.
- Clear representations of time-of-day features of spatial and temporal traffic patterns: the time-of-day traffic flow patterns can clearly identify the pattern changes over the same time period on different days or different time periods in the same weekday or weekend.

The spatial pattern identification presents meaningful results of recurrent traffic flow patterns: Over a certain time period, the heat map of spatial pattern anomaly in the region shows almost identical results on weekday and weekend. Two most abnormal intersections in the plots show that if one detector suffers a higher anomaly degree, other detectors in the same intersection are more probable to be abnormal. Further studies on sample locations show that the spatial traffic patterns remain almost the same for different days.
But if we focus on the traffic patterns in different time-of-day on the same day, an identical trend from AM peak, Noon, and PM peak can be found in different sub-regions.

The temporal traffic pattern identification demonstrates the non-recurrent features of traffic patterns. In our study area, the occurrence of temporal abnormal places are quite random, and detectors in the same intersections may have quite different anomaly degrees from each other. As compared to the spatial traffic pattern, temporal anomaly degrees in the same time period on different days usually have larger fluctuations and the higher value may indicate a bad traffic performance.

Also, it is proved that there exists almost no correlation between spatial and temporal anomaly degrees. The combination of the two leads us to a better understanding of traffic flow pattern. Results of a case study show the changes in the traffic flow patterns that are influenced by major municipal constructions over a certain scale of geography. The changes of land use and travel modes can, in turn, be identified as a shift in traffic flow patterns. By exploring the physical meanings behind the abnormal traffic flow patterns and comparing it with other state-of-the-art methods, one can see that the compression-based method identifies the abnormal points from a different perspective.

Although the proposed compression-based method is promising, its applications in transportation data analysis still show certain limitations. First, the method works better on a large dataset. It’s better to employ the method on the historical traffic data with long recording period. When dealing with small datasets, the results may not be convincing. Second, the method is very sensitive to outliers, and its application relies a lot on the credibility of the raw database. The researchers need to carefully prune the raw data before
generating anomaly degrees. Besides these shortcomings, this method provides insight into the interpretation of the multi-dimensional traffic-related data, and some research questions remain, especially in incident detection, network evaluation, and traffic management. Also, as the geographic distribution of spatial traffic flow patterns remains almost the same in different time periods, there may exist room for improving planning or management of the road networks. Strategies for improving the traffic jams can be evaluated by comparing the before and after anomaly degrees. The temporal traffic flow patterns indicate a traffic anomaly over a time period, and the causes of these anomalies are worth exploring in the future. Also, it is worth mentioning that we only include 5 categories of traffic-related data. It takes less than one minute for computing regional spatial traffic flow patterns over a 15-minute interval on a PC with intel i7 quad-core 3.7GHZ CPU and 32G RAM. We can improve the accuracy of the identification results of both spatial and temporal pattern identification by including more data categories such as link speed, queue length, etc. If so, the corresponding computational speed may be influenced and one can refer to Akoglu et al. (2012) for fast computation.
CHAPTER 4 EXPLORING TRAVEL BEHAVIOR: ABNORMAL MOVEMENTS USING HIGH-RESOLUTION TWEET TRAJECTORY

4.1 Empirical findings of tweet trajectory

4.1.1 Location features

The geo-tagged Twitter data contains four major attributes: user ID, time, location and current interests (Tweet content). The location is the combination of the latitude and longitude data whose precision can be as high as 0.001, which means the resolution of tweet trajectory is maximal 100 meters. The time and location information are paired and can jointly reveal the spatial features of the human mobility over different time periods in different days. One limitation of social media data is its representativeness (Chen et al., 2016). The tweeting frequency vary greatly among users and the enthusiastic users can tweet averagely 100 times per day. In this chapter, we target on around 200 tweeting users with high tweeting frequency and construct the hourly locations of these users in different days. For those users that tweet more than once during an hour, we take the distribution mode of these locations as the representative of the hourly location.

The location features can be interpreted in two aspects, shown in Figure 4-1 and Figure4-2. First, one can see in Figure 4-1 (a) that the most frequent visited locations take a large portion of the total locations over the same time period in different days. The ratios of most frequently visited locations show that despite different preferences, individuals prefer to stay within certain areas with a high probability. During daytime, more human activities possibly arise and thus lead to more travel. Therefore, the ratio of the most frequent visited location in the sleeping time is always higher than that in the daytime. Such
“location-focus” feature is observed for all travelers although there lie minor differences which is shown in Figure 4-1 (b). Some of the basic statistics are shown in Table 4-1.

Table 4-1 Basic statistics for location features of nearly 200 unique tweet users

<table>
<thead>
<tr>
<th>Number of tweets per traveler</th>
<th>Number of locations visited per traveler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>15408</td>
</tr>
<tr>
<td>Min</td>
<td>301</td>
</tr>
<tr>
<td>Average</td>
<td>1254</td>
</tr>
<tr>
<td>Median</td>
<td>750</td>
</tr>
</tbody>
</table>

(a)

(b)
Figure 4-1 The ratio of the most frequent visited locations over all locations for 5 individuals over different hour periods; (b) The ratio of the top 10 most frequent visited locations over all locations for the same 5 individuals at 18:00 p.m. in different days.

Second, the locations within a certain spatial scale are more likely to be conditioned under the same purpose. Previous studies show that the individuals live and travel in different regions, yet each user can be assigned to a well-defined area, defined by home and workplace, where she or he can be found most of the time (Gonzalez et al., 2008). The problem lies in how large the “well-defined area” should be. Figure 4-2 shows locations of the same 5 individuals as in Figure 4-1 over the same time period and these locations can be visually clustered at some of the most frequent visited area. The spatial scales of these areas vary from individuals and should be further justified.
Figure 4-2 Tweet locations of 5 individuals at 18:00 p.m. in different weekdays and the spatial scale of the most frequent places are labeled with a dashed circle and the radius in km.

Given that the location differences are relatively small within a cluster, some of the locations for a specific traveler as shown in Figure 4-2 may be driven by the same travel motivation but not necessarily in exactly the same place. This is one of the movement features of undirected travel and needs to be further identified.

4.1.2 Movement features

Exploring the human movement is a key component in the travel behavior study. The movement features are determined by not only the travelers’ preference but also the
geography of the area. What’s more, the physical activity of the displacement (movement) is effectively intervened by community and street-scale urban design and land-use policies and practices (Heath et al., 2006).

As shown in Figure 4-2, tweet trajectories can capture both movements driven by the directed travel and the rest influenced by the undirected travel. Figure 4-3 gives a closer look at the subareas of the traveler’s locations and the movements between locations. The locations of the destinations in Figure 4-3 can be visually divided into 3 distinct clusters. Further, the Euclidean distance between locations within the same cluster is employed to exemplify the effects of undirected travel. As one can see, there is a location center in each cluster that assembles most of locations as discussed in Section 2.1 and locations within the same cluster, together with the center location, is probably driven by the same travel motivation.
Figure 4-3 3 Movements from 17:00 p.m. to 18:00 p.m. in different days: The triangle point is the origin while the dots are destinations.

Figure 4-3 also indicates that the scale of undirected travel within a cluster can be measured. The scale can be approximated by checking the Euclidean distance between all locations within a cluster. Therefore, we use the Euclidean distance as a qualified representation of the effect of undirected travel instead of the actual trajectory, because it can measure the deviations of sampling locations from the real motivation-driven locations.

As a result, we plot the Euclidean distance between all locations as shown in Figure 4-4. The percentile values of the distances increase following a stair-like shape and a stair step in indicates a travelling distance between locations that is most frequently experienced by
the travelers. In this chapter, we define these “frequent distance values” between tweet locations as the “**featured distance**”. The longer featured distance may be accounted by many reasons such as the daily commuting between clusters (or locations within the clusters) or other regular activities discussed in Section 1. The shorter featured distance may represent the effect of undirected travel due to that some locations assemble within a cluster in which the corresponding undirected travel may be quite short.

![Figure 4-4](image)

**Figure 4-4** The percentile values of the distances between all locations of two travelers.

Further empirical studies also show interesting results in the “lower step” as shown in Figure 4-5. One can see that only a few of these discrete distances have high frequency that can be characterized as the featured distance.
The frequency of the distances between all locations of two travelers

Before quantitative analysis, the featured distance needs further study: First, every traveler has a different set of featured distance values and they should be quantified through massive data analysis instead of empirical determination; Second, for each traveler, locations within featured distance around the center location are not necessarily driven by the same travel motivation. This is true when two center locations are close to each other; For those locations between two center locations, we do further studies to judge which cluster they belong to; Third, our empirical results show that all travelers have a featured distance equal to 0.001 with highest probability. It means that the effect of undirected travel is probably to be scaled within a 100 meter.

4.1.3 Clustering features

As the locations in the same cluster are driven by the same motivation, these locations may follow a specific geo-distribution. Gonzalez et al. (2008) model this kind of geo-distribution as $\Phi_a(x, y)$, which is the probability to find an individual $a$ in a given position.
$(x, y)$. The distribution can be approximated by a normal distribution. We also validate similar results with high-resolution tweet location data.

Unlike the traditional probability distribution, the distribution of locations shows the normality feature of the location visiting times over a 2-dimension geographic span. After our exploration, each cluster shows a shape of multivariate normal distribution where the probability is approximated as the ratio of the visiting times of locations for a single traveler. Figure 4-6 (a) depicts the probability distribution within the cluster, while Figure 4-6 (b) shows the heat map of locations.

In each cluster, the center is visited most frequently and the shape of the distribution within the cluster may also be skewed due to the geography around the location center and consequently the different spatial scales of undirected travel. Even though, the normality features can be employed to build a viable clustering method. The featured distance discussed in Section 2.2 defines the possible boundaries of the cluster and we can employ the probabilities of locations in the distribution to determine whether they are driven by the same motivations as the center location.
Figure 4-6 the distribution of visiting times of all locations. The zero value of x-axis is fixed at the most frequent visited places and the x-axis values indicate the Euclidean distance of all locations from it. Positive values indicate the locations are on its east while the negative ones on the west. The y-axis values are normalized probability for the number of visiting times; (b) the heat map of visiting times of locations for the traveler. The location of plot (a) is circled.

4.2 Geo-Mobility method

As discussed in Section 2, the human travel locations can be grouped into several clusters which are confined within particular spatial scales and the locations in each cluster can be approximated by a normal distribution. To uncover travel behavior features, this chapter proposes a data-driven clustering method called Geo-Mobility method. This clustering method consists of 2 major stages: Geo clustering stage that properly groups the locations by the same travel motivation and Geo detection stage that identifies the trip purposes behind locations that result from abnormal movements.

4.2.1 Geo clustering

In the first stage, we extract the hourly geo-location data from the tweet trajectory. As the same to that introduced in Section 2.1, the hourly geo-location is represented by the one where the traveler tweets at the most times during the hour. Then we calculate the Euclidean distances between all hourly geo-locations during the observation period. These distances will be used to find the featured distance as defined in Section 2.2. We employ the X-means clustering method (Pelleg and Moore, 2000) to cluster the distance data and find the featured distance. The X-means method is built on the K-means clustering method and it can efficiently searches the space of cluster locations and number of clusters to
optimize the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) measure (Pelleg and Moore, 2000). We devise Algorithm 1 that can generally find the featured distance for each user as follows:

Algorithm 1: Featured distance clustering

**Input:** location coordinates (latitude and longitude information) of each traveler: \( C \).

**Output:** featured distance \( D_f \).

- extract from \( C \), the most frequent visited coordinates during each hour period in each day as the hourly geo-locations \( L^{dh} \): \( d \) is the day, \( h \) is the hour;
- calculate the Euclidean distance between all \( L^{dh} \): \( D = dist(L^{dh}) \);
- filter \( D \) by deleting all \( D = 0 \) and \( D \leq 0.01 \);
- implement X-means on \( D \) and obtain the featured distance \( D_f = \{D_f^i\} \).

In our study, we set \( D \leq 0.01 \) (about 1000 meters equivalent in NOVA) and one may set a higher value if needed. In the second stage, we cluster the locations of the travelers based on the featured distance. We start by calculating the frequency of each location during our observation period and iterate from the smallest \( D_f^i \). During each iteration, the most frequent location is selected as the center location. As the locations within the same cluster can be approximated by a normal distribution, we calculate the location probabilities within the range of \( D_f^i \) and justify whether they belong to the same cluster as that of the center location. The algorithm is as follows:

Algorithm 2: Probability-based location clustering

**Input:** featured distance: \( D_f = \{D_f^i\} \);
- hourly geo-locations for each traveler: \( L^{dh} \);
- confidence level: \( \alpha \);

**Output:** cluster results for \( L^{dh} \) of each traveler.

- define the unique geo-locations: \( U = unique(L^{dh}) \);
frequency of $U$ in $L^{dh}$: $F = \text{Freq}(L^{dh})$;
unclustered geo-locations: $U^0 = U$;

let $Num = 1$;
sort $\{D_i^f\}$ from the smallest to the largest;

\textbf{for} all $D_i^f$ in $D^f$ \textbf{begin}
select from $U^0$ a center location $U^c$ whose frequency $F^c = \max(F)$;
while sizeof $(U^0)>0$ \textbf{begin}
calculate $M$: the distance between $U^c$ and $U^0$;
calculate the threshold frequency for $j$th $U_j^0$ in $U^0$:
$$F_j^0 = F^c \cdot \exp\left(-\frac{1}{2 \frac{M_j \Phi^{-1}(1-2\alpha)}{D_i^f}}\right)$$, where $\Phi$ is Laplace Integral;
select all locations $T = \{U_j^0 | F_j \leq F_j^0, M_j \leq D_i^f\}$;
if sizeof $(T)>0$ \textbf{begin}
label $T$ and $U^c$ with Cluster ID: $C^i(T) = C^i(U^c) = Num$;
$Num = Num + 1$;
remove $T$ and $U^c$ from $U^0$;
else
label $k$th $U_k^0$ in $U^0$ with Cluster ID: $C^i(U_k^0) = Num + k$;
remove all locations from $U^0$ and $U^0 = \{}$;
end
end
$U^0 = U$;
end

\textbf{for} $U$ with more than one Cluster ID: $C^i$,
select $C(U) = \{C^i(U) | \arg\min_i (\exp(-\frac{1}{2 \frac{L_i \Phi^{-1}(1-2\alpha)}{D_i^f}})) \}$
label $L^{dh}$ according to $C(U)$.

Two major issues of Algorithm 2 are clarified as follows:

First: The qualification of a location to be included into a cluster is determined through a probability-based method by the visiting frequency at this location and the frequency at
its corresponding center location. For each center location, the clustering algorithm starts from the smallest featured distance. For the center location $C$, the visiting times (frequency) of its surrounding locations are assumed to be normally distributed across the geographic span as discussed in Section 2.3. The calculation of threshold frequency $F^0$ can be detailed in the following manner:

- Given a normal probability density distribution with mean equal to 0 and undetermined variance, the center location $C$ should be placed where $x = 0$ while its surrounding locations be where $x = M_j$ ($M_j$ is the distance between $C$ and $U^0_j$ as stated in Algorithm 2). The probability values of center location $C$ is $Prob(x = 0)$ while its $j$th surrounding location is $Prob(x = M_j)$. The features of frequency can be proved to be:

\[
\frac{Prob(x = M_j)}{Prob(x = 0)} = \frac{F^0_j}{F^c} \tag{4-1}
\]

Thus, $F^c \cdot \frac{Prob(x=M_j)}{Prob(x=0)}$ is the threshold frequency. For those lower than this value can be put into the same cluster as the center location. For those locations with $F$ higher than this threshold may be either the candidate of another center location or put into other clusters.

The cumulative density values with confidence level should be:

\[
\frac{1}{2} \left( 1 + \Phi \left( \frac{x - \mu}{\sigma \sqrt{2}} \right) \right) = 1 - \alpha \tag{4-2}
\]

When $\mu = 0$ and $x = M_j$, then:

\[
\sigma = \frac{M_j}{\sqrt{2} \Phi^{-1}(1 - 2\alpha)} \tag{4-3}
\]

The probability density function:
\[
Prob(x = 0) = \frac{1}{\sigma \sqrt{2\pi}}
\]  
(4-4)

\[
Prob(x = M_j) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(- \frac{M_j^2}{2\sigma^2}\right)
\]  
(4-5)

\[
F_j^0 = F^c \cdot \frac{Prob(x = M_j)}{Prob(x = 0)} = F^c \cdot \exp\left(- \frac{1}{2} \left[ \frac{M_j \Phi^{-1}(1 - 2\alpha)}{D_i^f} \right]^2 \right)
\]  
(4-6)

Second: A location belongs to the cluster in which \( F_j/F_j^0 \) is the minimum when there are more than one qualified center locations. We should choose the center location with highest \( F_j^0 \) assuming that there exist higher possibilities for this center location. Also, it is worth mentioning that this case is really rare.

\[
C(U) = \{ C^i(U) \mid \arg\min_i \left( \frac{F_j}{F_j^0} \right) \}
\]

\[
C(U) = \{ C^i(U) \mid \arg\min_i \left( \exp\left(- \frac{1}{2} \left[ \frac{M_j \Phi^{-1}(1 - 2\alpha)}{D_i^f} \right]^2 \right) \right) \}
\]  
(4-7)

One can see that our clustering method fully considers the unique features of tweet trajectory and groups time-of-day locations driven by the same travel motivation. As compared with popular density-based clustering method such as DBSCAN (Ester et al., 1996), the proposed probability-based method consider not only the density but also the possible distribution properties of these locations.

It is worth mentioning that the range of activity for a single traveler can be as large as an entire county. Thousands of hourly locations (The resolution of the each location can be as high as 0.01 in both longitude and latitude) can be captured by the tweet trajectory data and a traveler is able to travel to hundreds of locations during a whole year. Provided with the high-resolution tweet trajectory data, we can study the human mobility features within
this 100-meter range. Those locations that result from the abnormal movements may be driven by irregular social activities which can be traced in tweet contents.

4.2.2 Abnormal movement detection

We can unveil the travel behavior features by examining the human movements between different clusters because locations within the same cluster during the same hour period can be driven by the same purpose. In different days, during the same two continuous hour periods, the traveler may start traveling from the same origin cluster but end in different destination clusters. Weekday and weekend of the movements between two hour periods are intentionally separated. Thus, the ratios of each destination cluster in different days from the same origin cluster can be calculated. For those destination clusters with small ratio values should be taken as the irregular destinations. Here, we define the irregular destination locations as “abnormal locations” and the corresponding irregular directed travel between location clusters as “abnormal movement”. The algorithm is depicted as follows:

Algorithm 3: Abnormal movement extraction

Input: hourly location clusters for each traveler in different days: \( C^{dh} \);

percentage threshold: \( \beta^T \);

hourly geo-locations for each traveler: \( L^{dh} \);

Output: abnormal destination locations \( A^{dh} \).

\begin{algorithm}
\begin{algorithmic}
\State \textbf{for} each hour period \( h_i \) \textbf{ begin} \\
\hspace{1em} \textbf{for} location clusters \( C^{dh_i} = \{C^{dh_i}_j\} \) \textbf{ begin} \\
\end{algorithmic}
\end{algorithm}
calculate the percentages \( \{\beta^d_{ki}\} \) of all \( \{C^d_{ki}\} \) during the next hour period corresponding to the same \( C^{d(h_i-1)}_j \);

extract from \( L^d_{hi} \): the abnormal locations \( A^d_{hi} \) which is in \( \{C^d_{ki}|\beta^d_{ki} < \beta^T\} \);

end

end

The percentage threshold: \( \beta^T \) is set to be 0.05 in our study. Through Algorithm 3, the abnormal locations can be extracted and the corresponding tweets can be found consequently. Note that we extract all the tweet locations within the hour period: \( L^{d(h_i+1)} \) instead of the unique locations \( U \) as more tweets are involved to explain the travel motivation.

Figure 4-7 presents an example of a traveler’s locations from 18:00 p.m. to 19:00 p.m. and parts of the tweets posted at the abnormal locations. Figure 4-7 (a) shows most frequent visited location during 18:00 p.m. and Figure 4-7 (b) shows its corresponding destinations during 19:00 p.m. in different days. For display purposes, the cluster of destination locations are set in different colors and shapes.
Figure 4-7 location changes from (a) 18:00 p.m. to (b) 19:00 p.m. and the corresponding tweets: (1): “i hate this haircut”; (2): “ive never seen this movie that they playing on bet”; (3): “xisthatnigga marchmadness this is marchmadness this happening to duke makes this so great”.

We only choose three abnormal places for display purpose. People go to these places over a certain time period instead of the regular ones. The first two tweets are self-explanatory. For the third tweet for abnormal locations in Figure 4-7 (b), we can find the activity-related keywords: “marchmadness” which indicates “National Collegiate Athletic Association (NCAA) College Basketball”. Therefore, major social activities can be captured manually from tweets that provide a good explanation for some of the irregular trips.
4.3 Travel motivation decoding

The travel motivation of abnormal movement over a certain time period $h$ and in a certain day $d$ can possibly be decoded by checking the tweets the traveler posted in the corresponding locations $L^{dh}$. The traces of related social activities that drive the traveler making these movements can be found in social media data, such as tweets in this chapter. One limitation of this dissertation is that we do not have ground truth to justify whether a single tweet is related to the traveler’s actual activities or other special circumstances (Chen et al., 2016). In this sense, studies with further validation investigation are needed.

Compared to other information diffusion channel, Twitter created an online environment where content is created, consumed, promoted, distributed, discovered or shared for purposes that are primarily related to communities and social activities, rather than functional task-oriented objectives (Gal-Tzur et al., 2014). This dissertation will utilize tweets that are posted not only in the abnormal location $L^{dh}$ but also all tweets during that hour period $h$. Due to the complexity of travel behavior, the explanations behind these abnormal locations are diverse and some of them may not be easily assessable to the readers. Thus, we usually need a manual examinations and it is not a trivial task to explain all of them. Our interest lies in those activities that are driven by some major social events. This chapter particularly examines those events that can be summarized by some activity keywords or hashtags. For instance, in Feb. 13 2014, there was a keyword surge of “capitalweather”. People tweeted the delay caused by the “Biggest snow storm since Snowmageddon”. Public service info feed “Metrobus Info” lively broadcasted the congestion at North Capitol & New York Ave in DC area. This was widely retweeted by other users. The congestion in public transportation services echoes the findings in (Cools
et al., 2010; Lin et al., 2014) that weather conditions can change the travel behavior. One can see that the major social events or activities can arouse the interests of the tweet users resulting into the apparent pervasion of corresponding keywords over certain days. However, depending on the impact of the event, its related tweets will not typically last very long and usually disappear quickly in a short period.

Based on this feature, we improve the tweet activity decoding algorithm based on a keyword-search approach developed in (Zhang et al., 2016e). We first examine the total tweet collections throughout Year 2014 and search for the words that appear frequently in each day. Then, for all the frequent words, we select the words that appear frequently in one day but not so frequently or even vanish in other days; we call these words the activity keywords. Next, we manually interpret these words and the social activities behind them. Finally, we validate whether or not the tweets during the hour period of the abnormal locations contain the activity keywords.

**Algorithm 4: Tweet activity decoding**

**Input:** Tweet collections throughout Year 2014: $T^{dh}$; Tweets in the abnormal locations $A^{dh}$: $T^{adh}$; Word frequency threshold: $\gamma$; Word count threshold: $\tau$;  

**Output:** Activity keywords of each day: $K(d)$; Tweets of the abnormal locations contain those words: $T^{akdh}$;

for each $d_i$ begin  

decompose the tweets $T^{adh}$ into vectors of words $w(d_i)$;  

count frequency $f_{w(d_i)}$ of $w(d_i)$;  

eject the words:  

$W(d_i) = \{w(d_i) \mid f_{w(d_i)} \geq \gamma \sum f_{w(d_i)}\}$;
The setting of $\tau$ should not be either too large or too small. Large $\tau$ means that the social activities may last long, whereas in most of the case, travelers do not move more than once for irregular social activities such as a conference meeting or a festival parade. In our study, we set $\epsilon$ and $\tau$ to be 10% and 3 respectively. Through our examination, 46.2% of tweets posted in the abnormal locations contain the activity keywords and these abnormal locations can possibly be explained by the corresponding activities. Considering the capricious and unpredictable nature of the tweet contents and the coverage of Twitter, this percentage is high enough to constitute a good information source to explain the human mobility under irregular directed travel. Parts of the activities of abnormal movements are shown in Table 4-2. Given the word limits, we only list some of the keywords on February. To better understand these social activities and their impacts on the travel behavior surely involves more complex research projects in the future.

Table 4-2 Part of the keyword and the corresponding social activities

<table>
<thead>
<tr>
<th>Date</th>
<th>Keyword</th>
<th>Corresponding social activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/2/2014</td>
<td>bowl</td>
<td>Super Bowl XLVIII openday</td>
</tr>
<tr>
<td>2/2/2014</td>
<td>seahawks</td>
<td>ball game Seattle Seahawks vs. Denver Broncos</td>
</tr>
<tr>
<td>Date</td>
<td>Event</td>
<td>Details</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2/8/2014</td>
<td>openingceremony</td>
<td>2014 Winter Olympics Opening Ceremony</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>cancelled</td>
<td>Weather Alert on 2/13/2014. All Classes are Cancelled Today!</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>capitalweather</td>
<td>Biggest storm since Snowmageddon hits D.C.</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>patssnowstick</td>
<td>Pat's Snow Stick Challenge</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>pax</td>
<td>Winter storm pax hit middletown</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>shovel</td>
<td>Big snow</td>
</tr>
<tr>
<td>2/14/2014</td>
<td>chocolate</td>
<td>Valentine's Day</td>
</tr>
<tr>
<td>2/15/2014</td>
<td>aaroncarter</td>
<td>Washington DC private dinner &amp; listening party with Aaron Carter: an American singer</td>
</tr>
<tr>
<td>2/15/2014</td>
<td>oshie</td>
<td>Olympic: Oshie's shootout heroics power U.S. past Russia</td>
</tr>
<tr>
<td>2/15/2014</td>
<td>teamusa</td>
<td>Team USA Beats Russia In 'Marathon On Ice'</td>
</tr>
<tr>
<td>2/16/2014</td>
<td>beal</td>
<td>Basketball, Bradley Beal started the Wizards participation in the NBA</td>
</tr>
<tr>
<td>2/16/2014</td>
<td>dunk</td>
<td>Kevin Blaser Dunk vs. Lakewood Ranch Feb 16 2014</td>
</tr>
<tr>
<td>2/24/2014</td>
<td>tryouts</td>
<td>High School Spring Sports Tryouts 2014</td>
</tr>
<tr>
<td>2/28/2014</td>
<td>dance</td>
<td>KAR dance competition</td>
</tr>
<tr>
<td>2/28/2014</td>
<td>xpotomac14</td>
<td>xPotomac conference in Feb. 28 2014</td>
</tr>
</tbody>
</table>

### 4.4 Conclusions and discussions

This chapter proposes a social media based method to study the abnormal movements, i.e. irregular directed travel with high-resolution tweet trajectory. The involvement of Twitter in the study can bring a more detailed way to explore and explain the human mobility. A probability-based Geo-mobility method is developed to cluster the locations
and detect the abnormal locations. And a travel motivation decoding method is devised to explain these abnormal movements by identifying the social activities from tweets.

Several important empirical findings from tweet locations are summarized as follows:

- The travelers intend to stay within a certain range with a high probability. The travelers’ locations may assemble within several different clusters of different scales.

- Locations show unique clustering features and spatial location differences within a cluster can be the results of the undirected travel.

- In each location cluster for a single traveler, the visiting times of locations approximately follow a multivariate normal distribution across the geographic span.

To overcome the effects of the undirected travel, we propose the Geo-mobility method that groups the tweet locations into several different clusters driven by the same travel motivation. Abnormal locations and movements can be then detected accordingly. The movement between different clusters during different time periods and days are studied and less frequently visited locations are labeled as the abnormal locations.

The tweet posts are further examined to explore the travel behavior of abnormal movement. The possible explanations behind these locations can be found by the “travel motivation decoding” algorithm. 46.2% of these tweets posted in the abnormal locations contain keywords indicating major social activities.

One limitation of this chapter is that we lack ground truth of people’s travel motivation or trip purpose. Future studies shall take this limitation into account by conducting
supervised survey and experiments. In addition, an accurate automatic classification method for the social media data is worthy further investigations. With increasing coverage of social media, the automatic analysis of tweets can detect and even evaluate big social events and further their impacts on the travel behavior.
CHAPTER 5 THE HUMAN MOBILITY PATTERN STUDY BASED ON SOCIAL MEDIA

5.1 Human mobility study

5.1.1 Twitter displacement and human mobility patterns

In this chapter, we take each tweet user $u$ as a reliable data source and his tweets as a spatial-temporal text database: $U = [U_1, U_2, ..., U_l, ...]$. $U_i$ represents $i$th text entry in the database: $U_i = [D_i, L_i, T_i]$ where $D_i$ is the tweet content; $L_i$ is the location information of latitude and longitude, together with the label the CDP; $T_i$ is the time. The displacement between two time-sequential locations can be calculated as: $|L_i, L_j|$, where $| |$ refers to the function to calculate the Euclidean distance. Given an $U_i$, its corresponding sequential displacements over a certain time period can be calculated as shown in Equation (1):

$$S_i^W = [|L_i, L_j|: W^E \geq T_j - T_i \geq W^S] \quad (5-1)$$

Where $W^E$ and $W^S$ are the ending and starting time of the time window $W$. $L_i$ and $T_j$ refer to all the locations and timestamps of the tweet user. Thus, $S_i^W$ can be taken as domain of all displacements which start from $L_i$ and end after a certain time interval between $W^E$ and $W^S$; and all the tweet contents that are posted during these displacements can be notated in Equation (2):

$$D_i^W = [D_j: W^E \geq T_j - T_i \geq W^S] \quad (5-2)$$

The longest displacement of $S_i^W$ during $W$ can, to some degree, be taken as the biggest activity range for $L_i$; also, all corresponding tweet contents $D_i^W$ record the tweet user’s interests during $W$ and will be further studied to unveil the motivations of that displacement $\text{max}(S_i)$. This is mainly due to that people may tweet either before or after
when an activity starts and we can find the travel motivations with high probabilities during short time windows $W$. Thus, our study mainly selected three time windows: [0, 3600s], [3600s, 7200s], [7200s, 10800s]. The effects of different $W$ will be fully discussed and displacements over a 3-hour interval will not be included in the chapter.

For each $U_i$, its longest displacement during $W$ can be notated as the featured displacement of $U_i$: $S_i^W = \max(S_i^W)$ and the domain of $S_i^W$ is notated as $S^W = dom(S_i^W)$ which is the collection of all featured displacements of the tweet user. The corresponding $U_i$ of a featured displacement can be notated as $U_i^W = [D_i^W, L_i^W, T_i^W]$ and the domain of $U_i^W$ is noted as $U^W = dom(U_i^W)$. Here we define the mobility patterns of a tweet user as $M^W = domain(M_i^W)$ where:

$$M_i^W = [S_i^W, L_i^W, T_i^W, T_i^W]$$ (5-3)

From Equation (3), the mobility pattern of each user can be characterized by trip displacements, times and locations. According to our methods, there may exist more than one $U_i$ corresponding to the same $U_i^C$ as tweet users may tweet more than one time during a time window $W$. Thus, for a set of $M_i^W$ that has the same $I_i^W$, we only keep one $M_i^W$ whose $T_i^W - T_i$ is the largest. This filtering process can reserve all destinations of the trip displacements instead of all origins for further study. After this process, there will not be any two mobility patterns: $[M_i^W, M_j^W]$ in which $T_i \geq T_j$ and $T_i^W \leq T_j^W$. The mobility patterns can cover all the possible destinations of the tweet users together with its corresponding displacement over a certain time window. After filtering, we will not change the notation of filtered $M^W$ in this chapter for simplicity.

The Twitter displacements and mobility patterns derived from our method have certain advantages and worthiness as compared to that of previous studies.
• The geographic resolution of the displacements of Twitter can be much higher than that of bank notes (Brockmann et al., 2006), cellphone (Gonzalez et al., 2008), etc. The resolution can be as high as 100 meters which can well capture the intra-city and even intra-CDP movements. We can then study more public activities of interest with higher precision.

• The massive collections of Twitter data in NOVA also provide a good temporal resolution for displacement study. Our empirical examinations even find the consecutive hourly displacements for more than 400 tweet users. This will benefit the mobility pattern study as consecutive points have already been proved as useful in (Deng and Ji, 2010; Gonzalez et al., 2008).

• Our method assumes each Twitter location $L_i$ to be an underlying trip destination. Without fixed ranges of starting and ending time, the trip displacements can be employed to reveal the mobility patterns of the tweet users with higher flexibility.

However, the features of trip displacements $S^W$ based on Twitter data have drawn relatively less attentions in the previous studies. Even in some travel behavior literatures, the displacement features are usually overlooked and cannot be employed to infer the corresponding public activities due to lack of exact travelers’ location information. For example, the displacements in (Chen et al., 2016; Kang et al., 2012) may be inadequate as they are derived from inferred stops (e.g., between two cell power locations). Even though, most studies agree that combinations of multi-users’ displacements may comprise a dominant portion of short displacements together with several fragmentary portions of long displacements. The distribution of displacements of all tweet users in our study may
capture not only the uniformed decay characteristics of displacement frequency but also a population-based heterogeneity as shown in Figure 5-1.

Figure 5-1 (a) histogram of the displacements over 1-hour time window of all tweet users in this chapter; the time distribution of the displacements (b) lower than 100-m and (c) higher than 100,000-m; the day-of-week distribution of the displacements (d) lower than 100-m and (e) higher than 100,000-m.

Quite different from the studies based on cellphone or GPS data, we found an ever-dominant portion of displacements lower than 100-m, which accounts about for 74% of all displacement records and the portion of 0-m displacements even accounts about for 67%. The extremely unevenly distributed features can be explained in different ways: As the resolution of both latitude and longitude is 0.001, displacements lower than 100-m are not the typical displacements in our research scope. Under these conditions, tweet users may just have no special arrangements: browse and post tweets to kill time. Besides, unlike the
cellphones or bank notes, Twitter is sometimes an entertainment tool more than a necessary communication media and the Twitter records with no arrangements may be much more common. Consequently, Twitter can capture a extremely larger portions of 0-m displacements as compared to other data sources. Figure 5-1 (b) also provides side-proof for this showing that for the short displacements, nearly half of them are made after midnight and a very small portion be done in the morning which indicate that people are less likely to make a trip after midnight or in the morning. As compared, for the long displacements shown in Figure 5-1 (c), the portions for morning and midnight shrink to less than 25%.

Assuming that the heterogeneity inherent in the tweet users can partially be neutralized by the large amounts of tweet users involved, there are still concerns about the disparity of Twitter displacement as most people are more likely to tweet more activities when they have spare time as compared to when they are busy in case tweets potentially distract their concentrations. This may not be true as shown in Figure 5-1 (d) and (e) that there are nearly no weekday and weekend differences of tweet frequencies. Assuming that people typically have less activities on weekends, they still tweet as frequently as that in the weekdays. Above all, one can reasonably conclude that Twitter displacements have unique time and day distribution features and they can capture some fundamental mechanism driving human mobility patterns.

It is worth mentioning that displacements over 2 and 3-hour time windows also have the same features as shown in Figure 5-1. Our major interests lie in that long-distance displacements which, in most of the previous studies, can be extracted from the tail of a power law function with exponential cut-off (Clauset et al., 2009; Gonzalez et al., 2008):
\[ P(S_i^W) = (S_i^W + \gamma)^{-\beta} \cdot \exp(-S_i^W/\alpha) \] 

(5-4)

Where \( \alpha \) is taken as the cut-off values. Equation (4) suggests that the human mobility patterns (trajectories) can be approximated as a Lévy Flight whose step size can follow heavy-tailed probability distribution and this has been be proved in Figure 5-1 (a). Some displacement studies (Rhee et al., 2011) emphasize the truncated features of power law function because they believe it consists of two power functions separated at a conditioned value (Sjöberg et al., 2000). If so, the long-distance displacement, which is our major concern, should fall within the distribution in which \( S_i^W \) is larger than a conditioned value \( C \). Figure 5-2 shows that the truncated power law can fit the Twitter displacement well with two different conditional values. When \( C=100\)-m, \( \alpha \), \( \beta \) and \( \gamma \) are equal to 23833.9, 0.79 and 232.08; when \( C=2000\)-m, \( \alpha \), \( \beta \) and \( \gamma \) are equal to 6769.5, 0.4116 and 4456712.5, respectively. In both cases, the \( \alpha \) and \( \gamma \) values are different from most of previous studies because their displacement data have different magnitudes. However, the \( \beta \) values, which determines the fundamental mechanisms of the displacements, are quite smaller than 1.59 in (Brockmann et al., 2006) and 1.65 in (Gonzalez et al., 2008). Thus, the Twitter displacement may have different distribution shapes and thus record the human mobility patterns from different perspectives than that of bank notes, cellphone, etc.
Figure 5-2 Probability density value $P(S_i^W)$ of Twitter displacement with two conditioned values: $C=100$-m and $C=2000$-m.

5.1.2 Long-distance Twitter displacement

From the exemplary study of Twitter displacements with 1-hour time window in Section 3.1, we may find the extremely long displacements that are not routine in the mobility patterns over different time windows as shown in Figure 5-3 (a). One can see that Twitter displacements in each percentile only slightly increase when the corresponding time duration increases from 1 hour to 2 or 3 hours. This feature is even more apparent for those long-distance displacements framed in the rectangle. The maximum hourly displacement may be around 70 km which corresponds to 44 mph in movement. The results may indicate that the typical trip duration $T_{i}^{W} - T_{i}$ of displacements in NOVA may be less than one hour.
Figure 5-3 Distribution of Twitter displacement percentiles of (a) all tweet users over three different time windows; (b) each tweet user over time window [0, 3600s]; (c) each tweet user over time window [3600s, 7200s]; (d) each tweet user over time window [7200s, 10800s].

Besides the population-based features of the human mobility patterns, the mobility pattern for each user may display some kind of heterogeneity both in the scales but also the distributed features of the displacements. The framed parts in Figure 5-3 (b), (c) and (d) may be driven by special purposes of public activities which can be extracted from the
corresponding tweet contents $D_t$. The reasons why the long-distance displacements are important for the mobility pattern studies are that:

- These long-distance Twitter displacements, distinct from routine movements, are more likely to be driven by specific social events or personal interests. Detection of concentrations of public activities from Twitter is of great practical significance.
- The long-distance Twitter displacements are abnormal not only for the tweet users but also for the road networks. The mechanism of these activities may provide insights to the traffic management and control studies of derived travel demand forecasting and non-recurrent traffic congestions.

Thus, the long-distance displacements for each tweet user can be formulated as in Equation (5-5):

$$S_t^{LW} = [S_t^W : S_t^W \leq Q(S_t^W)] \tag{5-5}$$

Where $Q()$ calculates the percentile value of the displacements. The long-distance displacements vary from individuals, which can be as long as 10 miles to as short as 0.5 miles. This depends on the human mobility patterns of the individuals. One can reasonably assume $S_t^{LW}$ should start or end at some frequently visited places (e.g. home or workplace) as shown in Figure 5-4. The OD detection will not be covered in this dissertation. As compared to the whole study area, the displacement radius of gyration extends to a limited range within the whole year. Due to the page limits, we cannot show all the tweet users’ long-distance displacements. However, despite the different displacement magnitudes and geographic distributions, there are still some possible overlaps in the locations $I_t^{W}$. This makes possible the interpretation of the travel motivations from multi-users’ tweet contents.
It is worth mentioning that we employ the 95th percentile as a threshold in Figure 5-4, the effects of \( Q() \) will be further explored.

Figure 5-4 Geographic distribution of the origins and destinations of the long-distance Twitter displacements of two tweet users. The function \( Q() \) calculates the 95th percentile in both plots and the time window is set to be \([0, 3600s]\).

5.2 Travel motivation inference

5.2.1 Topic generation

For the long-distance displacements, the travel motivations can of the long-distance displacement can be inferred from their corresponding tweet contents. We can assume that each tweet can be comprised of a mixture of different topics and each topic can be represented as a bags of keywords. To generate the activity topics of long-displacement, for each entry in the mobility patterns \( M_i^W = [s_i^W, l_i^W, l_i^W, t_i^W, T_i^W] \), we mainly study:
• The tweet contents $D_i$ and $D_i^W$ which are posted on the origin and destination of the displacement, respectively on consideration that tweet users may post their interests at the starting and ending point of each trip.

• The tweet contents $D_i^C$ posted in the same CDP as that of $D_i^W$ during specific time period $[T_i^W - \Delta T, T_i^W + \Delta T]$ as other people nearby may tweet about the same public activities. Here the time interval $\Delta T$ for each $D_i^W$ is set to be 1 hour.

Thus, for each long-distance displacement, we can build a text database as shown in Equation (6):

$$R^i = [D_o, D_i^W, D_i^C]$$ (5-6)

Where $R^i$ is the text database for the $i$th mobility pattern and the $j$th text entry in $R^i$ is labeled as $R^i_j$. Then, we preprocess $R^i$ by three major steps: tokenization, stop-word filtering and stemming. In the tokenization process, the tweets contents are split by space and punctuation marks between each word, character, number or even Latinized symbol which are collectively called “token”. Meanwhile, a token may have different writing expressions in the tweets such as “accident” and its plural form “accidents” and deformed tokens with same root word should be taken equally. The deformation process is called stemming and we employ the Porter stemming algorithm (Porter, 1980) in this chapter. For these tokens, some of those may have an explicit meanings or indications while some others such as “I”, “by”, “the”, etc. do not. Those tokens have been summarized in (Hensher and Li, 2012) as stop-words and should be deleted from our text database. This process is called stop-word filtering which is prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). The process can be illustrated in Figure 5-5.
We can finally obtain a database with each entry containing a set of stemmed tokens without stop-words. For simplicity, we still label it as $R_j^i$. The topics, which are the abstract notions of activities, can be further derived from these tokens by the topic model: Latent Dirichlet allocation (LDA) (Blei et al., 2003). In this model, each stemmed token is the basic discrete unit:

$$R_j^i = [r_{j1}^i, r_{j2}^i, ..., r_{jk}^i, ..., r_{jN_j}^i]$$ (5-7)

Where $r_{jk}^i$ is the $k$th stemmed token in the $j$th text $R_j^i$; $N_j$ is the number of tokens in $R_j^i$.

According to Blei et al (2003), the process of LDA can be generalized as follows:

1) Assuming there are $L$ topics in $R^i$, we can choose a topic distribution $\theta_j \sim Dir(\alpha)$ in the $j$th text entry $R_j^i$; In each topic, we can choose a word distribution $\varphi_l \sim Dir(\beta)$ for the $k$th topic.

Where $\theta_j = [\theta_{j1}, \theta_{j2}, ..., \theta_{jL}]$ and $\varphi_l = [\varphi_{l1}, \varphi_{l2}, ..., \varphi_{lv}]$; $\varphi_{lv}$ represents a word proportion in the $k$th topic; $\theta_j$ and $\varphi_l$ both sums up to be 1;
(2) For each stemmed tokens \(r^i_{jk}\) in \(R^i\), we need to choose a topic \(z^i_{jk} \sim \text{Multinomial}(\theta_j)\) and a token \(r^i_{jk} \sim \text{Multinomial}(\phi_{z^i_{jk}})\).

Where \(z^i_{jk}\) is the topic identification of the token \(r^i_{jk}\).

\(\text{Dir}(\alpha)\) refers to the Dirichlet distribution with a set of unknown parameters \(\alpha = [\alpha_1, \alpha_2, ..., \alpha_L]\) which are real and positive. The function is shown in Equation (5-8):

\[
\text{Dir}(x|\alpha) = \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\prod_{i=1}^{L} \Gamma(\alpha_i)} \cdot \prod_{i=1}^{L} x_i^{\alpha_i-1}
\]

Equation (5-8)

The learning process to estimate the unknown parameters \(\alpha\) and \(\beta\) strictly follows that in (Blei et al., 2003). The results are a set of latent topics together with topic distribution in each text entry. One possible challenge lies in the LDA process is that it requires the pre-setting of the number of topics \(L\) which is also the dimensionality parameter in Equation (7).

There are a number of previous studies to find the appropriate number of topics but few of them have proved full adaptability in classifying the tweet contents. For the ensure the accuracy and correctness, we employ four different measures for the selection of \(L\). To our best knowledge, the most popular measure is proposed by Griffiths and Steyvers (2004) and determines \(L\) by the likelihood of each text entry conditioned by the number of topics. The second measure is proposed by (Cao et al., 2009) derived from the connection study between the best topic structure and the distances among topics. The third measure is proposed by (Arun et al., 2010) which computes the symmetric Kullback-Leibler divergence and finds that divergence values decreases significantly when the right number of topics. The fourth measure (Deveaud et al., 2014) propose simple heuristic that estimates the number of latent concepts by maximizing the information divergence between all pairs
LDA topics. As most of these studies validate their models and assumptions by long-text resources other than tweets, we averaged the ranks of the results from these different measures to choose the best number of topics.

The tokens in each topic may give a good description of the activities as shown in Table 5-1. Some may be the major public activities while others are just personal interests. Also, $D_i^W$ is the most important text entry in $R_i$ as compared to others because it is the tweet contents that are posted at the end of the displacement. Thus, we focus on the topic proportions in $D_i^W$. Besides this, as the word counts in tweets are usually less than 140 and typical tweet length may be 20-30 on average, there are only two topics for each text database $R_j^i$ under most circumstances. In this case, the topic words of first two topic portions in $D_i^W$ may be of the most significance.

Table 5-1 Examples of tokens and their corresponding public activities

<table>
<thead>
<tr>
<th>Date</th>
<th>Tokens</th>
<th>Public activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/14/2014</td>
<td>valentin</td>
<td>Valentine’s Day</td>
</tr>
<tr>
<td>7/4/2014</td>
<td>freedom</td>
<td>Celebration activity such as Freedom Fest 2014 fireworks</td>
</tr>
<tr>
<td>7/26/2014</td>
<td>silver</td>
<td>A new metro line: silverline, is opened</td>
</tr>
<tr>
<td>11/14/2014</td>
<td>cavali</td>
<td>Basketball games between Virginia Cavaliers and James Madison Dukes College</td>
</tr>
</tbody>
</table>

5.2.2 Topic tokens and their interpretations

Although the tokens may describe underlying public activities, the number of tokens in a topic may be so large that it is not easy to conclude essential descriptions for the activities. This is mainly due that most of tokens in a topic may not have so relatively weak event indications than the rest of minority. For example, some tokens such as “never”, “try”, “top” may appear more frequently than those such as “superbowl”, “freedom”, “wjmc2014”, etc.
because they are the former are more commonly used in our daily life. Thus, in this section, we further calculate the token frequencies in each topic and rank the corresponding words through a probability method. We notate the tokens in first two latent topics for the long-displacement $M_i^W$ as $U_i = [U_{i1}, U_{i2}]$. We assume that:

- A token in the topic should be more specific to their interests or special public activities if it is less mentioned by other people in the same day.

This assumption is only valid when there are large samples for study. We can extract all the tweets in the same day as $U_{i1}$ or $U_{i2}$, preprocess them as mentioned in Section 4.1 and calculate the frequency of the topic word in these tweets. According to the frequency of each token, we can rank the tokens as shown in Table 5-2. Two exemplary long-distance displacements are extracted by Equation (5) in which $Q()$ calculates the 90th percentage value. The tokens of the latent topics with lower frequencies mostly come from the specified vocabulary of the tweet user and may direct to some events. For instance, “swagdiqha”, “rashabokhari”, “sweezy32”, etc. refer to the names of tweet users; “jacoblong23” may refer to the basketball game between Deer Valley and Oakland Tech in which the former team wins 23 points. The topic words should be interpreted together with their time and space attributes and most of them can accessible to a public activity by careful interpretation. The difficulties lie in automatic relating the topic words to social events and the process usually needs manual examinations due to the great varieties in both the social activities and people’s interests driven by the unpredictable and capricious of human mind. It is worth mentioning that 410-m displacement is taken as long-distance in Table 5-2 just because most of the hourly displacements of this tweet user are equal to 0-m.
Table 5-2 Token rankings in two exemplary displacements of a tweet user over time window [0, 3600s]

<table>
<thead>
<tr>
<th>Displacement distance: 410 m</th>
<th>Displacement distance: 12915.96 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date: 04/15/2014</td>
<td>Date: 04/15/2014</td>
</tr>
<tr>
<td>Starting time: 00:04:34</td>
<td>Starting time: 17:27:49</td>
</tr>
<tr>
<td>Ending time: 00:34:42</td>
<td>Ending time: 18:07:05</td>
</tr>
<tr>
<td>Origin CDP: Germantown</td>
<td>Origin CDP: Rockville</td>
</tr>
<tr>
<td>Destination CDP: Germantown</td>
<td>Destination CDP: Germantown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 1</th>
<th>Topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>swaggdiaqha</td>
<td>musicdontmatch</td>
<td>rashabokhari</td>
<td>spici</td>
</tr>
<tr>
<td>0.000052</td>
<td>0.000052</td>
<td>0.000052</td>
<td>0.000052</td>
</tr>
<tr>
<td>wildddd</td>
<td>stxox</td>
<td>swaggdiaqha</td>
<td>plu</td>
</tr>
<tr>
<td>0.000052</td>
<td>0.000052</td>
<td>0.000052</td>
<td>0.000522</td>
</tr>
<tr>
<td>soooooo</td>
<td>yocynthiaaaaaa</td>
<td>sweezy32</td>
<td>easi</td>
</tr>
<tr>
<td>0.000104</td>
<td>0.000052</td>
<td>0.000157</td>
<td>0.00094</td>
</tr>
<tr>
<td>neighborhood</td>
<td>crank</td>
<td>jayciduwop</td>
<td>chicken</td>
</tr>
<tr>
<td>0.000365</td>
<td>0.000104</td>
<td>0.000209</td>
<td>0.001566</td>
</tr>
<tr>
<td>oculu</td>
<td>jacoblong23</td>
<td>wiz</td>
<td>without</td>
</tr>
<tr>
<td>0.000047</td>
<td>0.0000104</td>
<td>0.0000209</td>
<td>0.002558</td>
</tr>
<tr>
<td>bull</td>
<td>lifestyl</td>
<td>baeex</td>
<td>ll</td>
</tr>
<tr>
<td>0.0000679</td>
<td>0.000104</td>
<td>0.0000261</td>
<td>0.00355</td>
</tr>
<tr>
<td>neg</td>
<td>toronto</td>
<td>page</td>
<td></td>
</tr>
<tr>
<td>0.000731</td>
<td>0.000104</td>
<td>0.0000522</td>
<td></td>
</tr>
<tr>
<td>fat</td>
<td>musicthatdontmatch</td>
<td>plz</td>
<td></td>
</tr>
<tr>
<td>0.001514</td>
<td>0.000157</td>
<td>0.000626</td>
<td></td>
</tr>
<tr>
<td>ha</td>
<td>sweezy32</td>
<td>neg</td>
<td></td>
</tr>
<tr>
<td>0.001514</td>
<td>0.000157</td>
<td>0.000731</td>
<td></td>
</tr>
<tr>
<td>top</td>
<td>jayciduwop</td>
<td>stfu</td>
<td></td>
</tr>
<tr>
<td>0.001723</td>
<td>0.000209</td>
<td>0.001044</td>
<td></td>
</tr>
<tr>
<td>wrong</td>
<td>baeex</td>
<td>posit</td>
<td></td>
</tr>
<tr>
<td>0.003132</td>
<td>0.000261</td>
<td>0.001096</td>
<td></td>
</tr>
</tbody>
</table>

The time window in Table 5-2 captures travel motivations of the long-distance displacement in one hour time window. The results may capture most of the long-distance displacements and their destinations but cannot be intact without that of other time windows. Also, the selection of $Q()$ is also crucial to our study. Figure 5-6 shows that data of 1 hour time window capture most of the locations and their long-distance displacements in other two time windows and the ratio is higher in [7200s, 10800s] than that in [3600s, 7200s]. Higher values of time window $W$ may not be necessary because they may just capture the long-distance which are the combinations of those of other time windows. Also,
higher percentage setting may increase the ratio but also shrink the size of the displacement data.

According to Table 5-2, the topic tokens, especially those with lower frequency values, partially reveal the language habits of the tweet users. However, after careful examinations, we can see that some of them give a clear indications of public activities while some others may not. For the topic words without activity indications, there may be two possible explanations:

- These long-distance displacements $S^W_i$ may not be driven by activities. Instead, they are the typical undirected travel in which destination is sometimes ancillary to the travel rather than the converse which is usually assumed (Mokhtarian and Salomon, 2001)
- The tweet database $R^t_i$ in Equation (6) may not give a description of the activities. Instead, they may focus on other things such as describing self-mood or talking with friends.

Even though, the topic words may be a more reliable description of what people think, say or act during their displacements.
Figure 5-6 The long-distance displacement ratio of time window [0, 3600s] to [3600s, 7200s] and [7200s, 10800s] and its corresponding size with different percentile calculator $Q_O$.

### 5.3 Applications and discussions

In this section, we made two case studies to exemplify the applicability of our methods in unveiling both the individual mobility patterns and the influences of major social events. In both cases, we can interpret most of the topic words through web search and obtain the potential activities behind them combining with the corresponding time and location information. For each long-distance displacement, we will start searching from topic both from topic 1 and topic 2 as introduced in Section 4.3 by the order of word frequency. If we did not find any clues of activities till the 10th topic words, we can assume that these topic words do not correspond to a public activity which indicate that the tweets do not unveil the travel motivation of the long-distance displacements. Thus, interpreting all these topic
words through manual interpretation is an arduous task due to the large volumes and great variesties of the human activities.

5.3.1 Individual mobility pattern and travel motivation identification

In this case study, we focus on the individual mobility patterns of long-distance displacements and try decode their travel motivations. This is important because most of long-distance displacements are driven by the special travel motivations. We randomly select one tweet user for our exemplary study. His mobility patterns and travel motivations are studied and unveiled as described in Section 4.

Figure 5-7 and Table 5-3 jointly illustrate partial mobility patterns of long-distance displacements of the tweet user. Figure 5-7 (a) shows all of his long-distance displacements from May 1 to 15. The displacements shown are not consistent in time because people may not tweet within a regularly time period. Sometimes, they tweet when they start the trip but do not when they are back, or just reverse. Figure 5-7 (b) shows several longest displacements in the whole year. It also demonstrates his mobility radius in the urban road network. Both figures show clear centers which indicates the home or workplace.

Also, we cannot decode all the travel motivations of the displacements because of not only the difficulties of pinpointing a specific event by just a few topic words but also the lack of sufficient information. Even though, one may still find some interestingness the tweet user want to express during the displacements as shown in Table 5-3. The individual mobility pattern identification tells not only the places of interests that they travel to but also when they make the displacements, especially those of long-distance displacements.
Figure 5-7 Long-distance displacements (a) from 5/1/2014 to 5/15/2014, displacements on the same day are in the same color. (b) longer than 3000 m in the whole year. The
Table 5-3 The topic word interpretations of the long-distance displacements (a) Long-distance displacements from 5/1/2014 to 5/15/2014; (b) Long-distance displacements longer than 3000 m in the whole year

(a)

<table>
<thead>
<tr>
<th>TripID*</th>
<th>Start time</th>
<th>End time</th>
<th>Topic word interpretation</th>
<th>Distance (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5/2/2014 19:33</td>
<td>5/2/2014 20:04</td>
<td>Not agree to deactivate his twitter</td>
<td>206.14527</td>
</tr>
<tr>
<td>5</td>
<td>5/5/2014 15:21</td>
<td>5/5/2014 16:05</td>
<td>-*</td>
<td>618.47261</td>
</tr>
<tr>
<td>7</td>
<td>5/9/2014 0:22</td>
<td>5/9/2014 1:05</td>
<td>-*</td>
<td>3081.74538</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>TripID*</th>
<th>Start time</th>
<th>End time</th>
<th>Topic interpretation</th>
<th>Distance (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5/9/2014 0:22</td>
<td>5/9/2014 1:05</td>
<td>-*</td>
<td>3081.74538</td>
</tr>
<tr>
<td>5</td>
<td>8/12/2014 3:23</td>
<td>8/12/2014 4:06</td>
<td>Travel with his father</td>
<td>7934.7021</td>
</tr>
</tbody>
</table>

* The “TripID” corresponds to that in Figure 5-7.
* “-” indicates not enough information for interpretation.

5.3.2 Social event influence identification

Besides the individual pattern identification, the travel patterns of the people towards a special social event are also worth studying as they correspondingly can unveil the potential
traffic impact. We can start from extracting the displacements containing the topic words related to a certain event.

Our exemplary social events are the newly-built 11-mile extension of Silver Line, a subway line of Washington Metro starting service on July 26, 2014. The topic words we select include the names of the metro line and 5 new stations: “Silverline”, “Spring Hill”, “Greensboro”, “Tysons Corner”, “McLean”, and “Wiehle-Reston East”. Except “Silverline”, the rest topic words are proprietary not only to the metro stations but also to the commercial facilities around the Silver Line. As the topic words are the representations of the users’ interests, Figure 5-8 shows an increase of focus in the general public around July 26: the opening day of the metro line.

![Figure 5-8](image)

Figure 5-8 The counts of daily tweets that contain “Spring Hill”, “Greensboro”, “Tysons Corner”, “McLean”, “Wiehle-Reston East”, “Silverline”.
This “word surge” indicates that a major civil construction may arouse special interests and parts of these interests may even induce trips. Thus, one can reasonably assume that some of the displacements from July 26 ~ 28 are induced by the opening of Silver Line. As the origins and destinations of these displacements can give insights to the traffic management and control, inferring the locations of the activities is an interesting topic in some of the previous studies (Chen et al., 2014a; Ikawa et al., 2012). One can even find insightful results in our case study. We extract the long-distance displacements which contain their topic words: “Silverline”. One can see that most destinations of these displacements are around the locations of the Silver Line stations as shown in Figure 5-9 which proves the validity of our methods.

Figure 5-9 the locations where the “Silver Line” displacements end (destination) from July 26 to July 28.

Besides where-they-go, one can even find the origins of these displacements as shown in Figure 5-10. The Census-designated places (CDPs) of these displacements give an
approximate ratio of where these people come from. It will be of greater practical significance in future studies with the penetration of Twitter together with other social media tools.

Figure 5-10 the ratios of CDPs where the “Silver Line” displacements start (origin) from July 26 to July 28.

5.4 Conclusions and discussions

This chapter did a thorough study on the human mobility patterns, extract the abnormal Twitter displacements and further explore the explanations behind these displacements from the tweet contexts.

We take each tweet user $u$ as a reliable data source and his tweets as a spatial-temporal text database. We further define the mobility patterns of a tweet user as the combination of displacements, the locations of origins and destinations, starting time and ending time within a certain time window. Within 1-hour time window, our findings are quite different
from that based on cellphone or GPS data: an ever-dominant portion of displacements lower than 100-m were found which accounts about for 74% of all displacement records and the portion of 0-m displacements even accounts about for 67%. This may be due to that the resolution of Twitter is much higher than those datasets in the previous studies.

The Twitter displacements can be approximated by a Lévy Flight whose step size can follow heavy-tailed probability distribution. When the conditioned value equals to 100-m and 2000-m, $\beta$ value, which determines the fundamental mechanisms of the displacements, is 0.79 and 0.41, respectively, which are quite smaller than 1.59 in (Brockmann et al., 2006) and 1.65 in (Gonzalez et al., 2008). The extracted long-distance displacements with 95th percentile threshold start or end at some frequently visited places and the displacement radius of gyration extends to a limited range within the whole year.

We employed the Latent Dirichlet allocation (LDA) to extract the topics from the tweet contents behind these long-distance displacements. Our results show that the tokens of the latent topics with lower frequencies mostly come from the specified vocabulary of the tweet user and may direct to some events.

Two promising applications of our study are:

- First, for each individual, we can unveil his mobility patterns and decode his travel motivations behind the long-distance displacements. The individual mobility pattern identification tells not only the places of interests that they travel to but also when they make the displacements, especially those of long-distance displacements.
- Second, based on selected topic tokens, we can also study the travel patterns of the people towards a special social event. In our example, the word surge: “silverline”
indicates a major civil construction that arouse movements and it will be of great practical significance to study the origin of these displacements

Our study opens a window to access the human mobility features through Twitter displacement. It not only brings a mobility pattern result with higher resolution but also interpret the travel motivations behind the long-distance displacements. The current interpretation method may be biased because they are mainly manual labor. This is inevitable due to the large variety of human mobility activities. It can be more specific when we focus on smaller research scope. Further studies can be more accurate with increasing coverage of tweets as a social communication tool. In addition, in future attempts, we can even combine the traffic data and social media data to see the impact of long-distance displacements on the travel time, traffic flow through put, etc.
CHAPTER 6 TRAFFIC ACCIDENT DETECTION WITH BOTH SOCIAL MEDIA AND TRAFFIC DATA

6.1 Data preprocessing

6.1.1 Tweet data preprocessing

We preprocess the raw data of tweet data to constitute the database that can be used for further analysis. The first step is to extract the candidate tweets that possibly describe the on-site traffic accidents. Usually, these candidate tweets should contain one or more keywords such as “accident” or “crash” that are accident-related and we can assume that people describe traffic accidents by accident-related words. However, there has been no consensus on such a vocabulary of the accident-related words. Thus, we turn to the traditional news media and collect about 100 articles of news that discuss the traffic accident. In all these articles, we select the words that appear the most frequently. The frequency of a word is the times that a specific word appears in these articles. Except the common words such as “I”, “is”, etc. and those that reflect specific geographic and event features, we found that most of the articles contain a common list of words with a high frequency as shown in Table 6-1.

Table 6-1 Accident-related words


The second step is to extract the candidate tweets based on the accident-related words. We apply a filter based on keywords to obtain the accident-related tweets. As compared to traditional media, social media blogs are generated by the crowds without any editorial
review. Some of the words may have grammatical errors. Thus, to ensure both the accuracy and sample size, certain rules are further implemented:

- Include the words that are relevant to accidents but apparently misspelled or personally modified including “acident”, “incdent”, etc.
- Include other variations of accident-related words such as the word pairs that have a hyphen in word pair such as roll-over, etc.
- Exclude the tweets related to transportation authority or news media.

Finally, we obtained more than 3500 candidate tweets. These candidate tweets are later used to train the accident detection model.

6.1.2 Token filtering and stemming

To fetch the proper input features, each tweet is further decomposed into components. These components may including words, characters, numbers or even Latinized symbols that are collectively called “tokens”. There are more than 10000 tokens from all eligible tweets. We assume that some of the tokens may have no explicit meanings while some other tokens can potentially convey one or more instantaneous ideas and feelings of the tweeters. Part of them will be selected as the features after necessary filtering and stemming. The steps can be illustrated in Figure 6-1.
First, the punctuation marks convey almost no meanings and should be discarded, and all other words should be converted into lower case. Meanwhile, some of the words or characters that have no apparent linguistic meanings or significant event indications should be filtered out before the processing. These words are referred as stop-words. Stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). The stop-word list we used refers to Ranks-NL (2015).

Second, some of the words have different writing expressions due to the grammatical reasons but convey almost the same meanings such as “accidents” and “accident”. The token stemming is necessary to reduce these inflected (or sometimes derived) words to their word stem, base or root form. In this chapter, we employ the Porter stemming algorithm (Porter, 1980) for the token stemming and each token is grouped into the proper stemmed token.

After token filtering and stemming, each tweet $T_i$ can be summarized with a set of stemmed tokens. Of all the tweets $T$, there are more than 3000 stemmed tokens symbolized.
as \{t_1, t_2, \ldots, t_j\}. The stemmed tokens are the features for each tweet \(T_i\) and each tweet has different token features. If the tweet contains a stemmed token, the corresponding token features are labeled as 1 otherwise 0. Thus, the token features and the tweets \(T\) form our binary database \(D_S\) and it will be used for the feature selection.

Token filtering and stemming is the initial and necessary step for regression analysis in different models.

6.2 Classification by SVMs

6.2.1 The entire process

This section employs the supervised learning model SVMs (Karatzoglou et al., 2005) and the process of inferring functions of these models. The supervised learning consists of two major components: labeling and modeling. Labelling refers to the manual labeling process on the candidate tweets and the manual label is the categorical value assigned to each tweet. In our study, the two-class manual label is employed deciding whether the tweets are accident-related or not. After labeling, more than 400 tweets are taken as accident related. To generate a balanced dataset, we randomly select non-accident-related tweets which are twice the size of accident-related tweets and combine them to constitute a tweet database. These tweets are symbolized as \(T = \{T_1, T_2, \ldots, T_i, \ldots, T_M\}\) and \(i\)th tweet is \(T_i\). The corresponding label for \(T_i\) is \(L_i\).

The model inputs include the features extracted from the tweet and the traffic information. The input features are one of the major concerns in this chapter and will be fully detailed later. SVMs can employ different kernel functions to keep the computational load reasonable. In our study, we employ the linear kernel to train and predict the models.
In the process of model training, we further implement 5-fold cross validation (Geisser, 1993) to increase the accuracy of the predicted model. Cross-validation can give insights on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data. We finally obtained an overall estimation by averaging 5 test results.

6.2.2 Classification with individual word

This section describes the steps of selecting features from individual token in the database $D_S$ and the corresponding results. We focus on correlation between the individual token and our manual label. The correlation benchmark we choose is phi coefficient (Cramér, 1999), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as \( \phi \)) between two variables \( x \) and \( y \) is calculated as:

\[
\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{11}n_{00}n_{10}n_{01}}} \tag{6-1}
\]

Where all notations are defined in the following table:

<table>
<thead>
<tr>
<th></th>
<th>( y = 1 )</th>
<th>( y = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 1 )</td>
<td>( n_{11} )</td>
<td>( n_{10} )</td>
<td>( n_{1*} )</td>
</tr>
<tr>
<td>( x = 0 )</td>
<td>( n_{01} )</td>
<td>( n_{00} )</td>
<td>( n_{0*} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{*1} )</td>
<td>( n_{*0} )</td>
<td>( n )</td>
</tr>
</tbody>
</table>
Those tokens whose $|\phi|$ is higher than 0.1 are selected. Following this rule, 27 tokens are selected and some of them are shown in Figure 6-2.

![Figure 6-2 Correlations between the manual label and the individual stemmed tokens](image)

From Figure 6-3, the stemmed tokens may be different from their original words in which “accid” refers to “accident”; “accident” does to “accidently”; “incid” does to “incident”. Some of the tokens may be accounted by the geographic uniqueness such as “66”, “95”, and “495” which indicates the route number, and this means the tweet intends to report the traffic accidents with route name; some may be the topic-related words.
including “traffic”, “accident”, etc.; other words such as “damage” or “accidentally” are too
general in our daily lives and thus lose the uniqueness in describing the traffic accident.

With selected individual tokens as the input of the regression model, one can simply
compare the results by token features. To evaluate the achieved results in different models,
we employed statistical metrics: accuracy and precision:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \\
\text{Precision} = \begin{cases} 
\frac{TP}{TP + FP} & \text{for accident-related} \\
\frac{TN}{TN + FN} & \text{for non-accident-related}
\end{cases}
\]

Where

<table>
<thead>
<tr>
<th>Groundtruth = 1</th>
<th>( P_{Model} &gt; 0.5 )</th>
<th>( P_{Model} \leq 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( TP )</td>
<td></td>
<td>( FP )</td>
</tr>
<tr>
<td>( FN )</td>
<td>( TN )</td>
<td></td>
</tr>
</tbody>
</table>

In Equation (3), precision is calculated for accident and non-accident, respectively. One
can set the threshold of different \( \phi \) to determine the number of the token features in the
regression model. Theoretically, selecting more token features increases the accuracy and
precision of the regression results but also increases the computational time, model
complexity, and even causes over-fitting as shown in Figure 6-3. When we set the threshold
of \( \phi \) as 0.2, there is only 4 qualified tokens. With the decreasing of the threshold, the
number of individual tokens increases as expected while the accuracy of the prediction also
increases slightly. When the threshold of \( \phi \) is equal to 0.15, we can obtain an accuracy of
0.784 with 11 individual tokens. With decreasing of \( \phi \), more tokens involved will not
improve the results significantly. One can see that with fewer than 15 words, the tweets
can be used to classify an accident with an accuracy around 0.78. However, this can be further improved.

![Figure 6-3](image)

**Figure 6-3** Regression results with selecting individual tokens under different thresholds of correlation coefficient $\phi$

### 6.2.3 Classification with paired words

Features from individual token are not be sufficient to automatically classify the tweets because these may overlook the associations between words and such associations can have much more significant indications than single ones. For example, in a tweet post, the occurrence of word “car” conditioned by “accident” may increase the accident-related probability. Conversely, the occurrence of token “car” conditioned by “maintenance” or “repair” may undermine the likelihood of accident-related tweets.

In this section, we select the features from paired words by studying the association rules between the manual label and the stemmed tokens in the binary database $D_S$. The association rules can be unveiled by the Apriori algorithm (Agrawal and Srikant, 1994;
Hahsler et al., 2007). Apriori algorithm finds the regularities in large-scale binary data by two major probabilities: support and confidence.

We label all stemmed tokens in the database as $t = \{t_1, t_2, ..., t_j, ..., t_N\}$. Given a stemmed token $t_j$, support of $t_j$ is the proportion of tweets which contains $t_j$ in the database.

$$
supp(t_j) = \frac{sizeof([T_i, t_j \subseteq T_i])}{sizeof([T])}
$$

(6-4)

Where $t_j$ is the $j$th token; $T_i$ is the $i$th tweet. Setting a threshold of $supp(t_j)$, we can filter out a limited number of qualified $t_j$. Similar to the support of each individual token, we can even calculate the support of paired tokens $supp(t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m})$:

$$
supp(t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m}) = \frac{sizeof([T_i, t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m} \subseteq T_i])}{sizeof([T])}
$$

(6-5)

Where $j_1 \neq j_2 \neq ... \neq j_m$. The paired tokens can be the combination of any two or more individual tokens. The concurrent tokens in one tweet post are quite common such as “traffic accident”, “severe injury”, etc. One can see that support deals mainly with the frequencies of one or more tokens. As the tweet database are filtered according to several different keywords, the word combinations of accident-related tweets may be also quite different. Thus, support of paired tokens can possibly capture different concurrent tokens that can possibly be used as the features in the model. But not all of them may be qualified as the features in the model. Besides support, the association rule between manual label and the paired tokens can be further revealed by confidence, defined as:

$$
conf(L_i \Rightarrow t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m}) = \frac{supp(L_i \cap t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m})}{supp(t_{j_1} \cap t_{j_2} \cap ... \cap t_{j_m})}
$$

(6-6)
Where $L_i$ represents the accident label. In the confidence calculation, we focus more on paired tokens that are related to traffic accident when $L_i$ is equal to 1 in Equation (6). The number of individual tokens in a paired token is always more than 1 and the maximum can theoretically be equal to the number of tokens in a tweet post. Also, if one increases the size of the paired tokens, the computational time will dramatically increase bringing almost no benefit. Our initial examinations show that almost no association rule exists in tweets when the size of paired tokens is larger than 4.

In most of the previous studies, setting support and confidence is sometimes mandatory. The setting of support can be a small value that can include as many as paired tokens for feature selection. The setting of confidence, as compared, usually influences the results significantly and different values should be further investigated in the classification for the impacts. We conduct an empirical study to see how the token features can reveal the language of customs of tweets in describing traffic accidents. When support is equal to 0.01 and confidence is equal to 0.1, our results show that most paired tokens contain “accident”. The list of all paired tokens is traced in the findings as shown in Table 6-2.

Table 6-2 Paired tokens selected by the Apriori algorithm

<table>
<thead>
<tr>
<th>accid*</th>
<th>vattraff</th>
<th>accid</th>
<th>close</th>
<th>accid</th>
<th>got</th>
</tr>
</thead>
<tbody>
<tr>
<td>accid</td>
<td>mile</td>
<td>accid</td>
<td>road</td>
<td>accid</td>
<td>im</td>
</tr>
<tr>
<td>accid</td>
<td>wtop</td>
<td>lane</td>
<td>block</td>
<td>accid</td>
<td>car</td>
</tr>
<tr>
<td>accid</td>
<td>major</td>
<td>accid</td>
<td>block</td>
<td>accid</td>
<td>lane</td>
</tr>
<tr>
<td>accid</td>
<td>left</td>
<td>lane</td>
<td>wtoptraff</td>
<td>accid</td>
<td>lane</td>
</tr>
<tr>
<td>accid</td>
<td>near</td>
<td>accid</td>
<td>wtoptraff</td>
<td>accid</td>
<td>car</td>
</tr>
<tr>
<td>accid</td>
<td>95</td>
<td>accid</td>
<td>bad</td>
<td>accid</td>
<td></td>
</tr>
<tr>
<td>accid</td>
<td>66</td>
<td>accid</td>
<td>lane</td>
<td>accid</td>
<td></td>
</tr>
<tr>
<td>accid</td>
<td>rd</td>
<td>accid</td>
<td>traffic</td>
<td>accid</td>
<td></td>
</tr>
</tbody>
</table>
accid involv | car got

*accid refers to the stemmed token of the word “accident”.

Table 6-2 shows that most of the association rules contain the token: “accid” (referred as “accident”) and most of other tokens can also be found in the results of correlation study as shown in Figure 6-2. Same as individual tokens in Section 3.3, the paired token features in the database are equal to 1 if the tweet contains the corresponding paired tokens and 0 otherwise. We perform another analysis by incorporating paired token features into the SVM model. Recall that when selecting individual tokens, the threshold of $\phi$ was set to be 0.15; when selecting paired tokens, support is set to be 0.01 to ensure that many possible paired tokens are included, whereas the confidence value is manually changed from 0.1 to 1.

![Graph showing regression results with paired tokens under different values of confidence where support is equal to 0.01](image)

Figure 6-4 Regression results with paired tokens under different values of confidence where support is equal to 0.01
As one can see from Figure 6-4, the major findings are that incorporating the paired token features improves the accuracy by 2~3%. By comparing the results of different confidence values, one can find that high value of confidence will not obtain enough paired tokens to improve the accuracy of the prediction. With confidence value decreasing, more paired tokens are involved in the model as expected but only a few of them can improve the accuracy and precision. Our analysis shows that when confidence is set to be 0.6, approximate 25 paired token features are selected, resulting in an accuracy of 0.808. Recall that around 100 individual token features equivalently obtain an accuracy around 0.8. Therefore, paired tokens are more efficient in using fewer words to obtain the same result.

It is worth mentioning that our study only focuses on the association rule between tokens and the manual label, but not that between all tokens. Thus, one can call it a supervised association rule mining, providing satisfactory results with much fewer features. The advantages of employing association rules in studying the accident-related tweets can be summarized as:

- As compared to the detailed accident reports or web news, the tweets are limited to 140 words and the word counts may be even less in practice. Tweet users are more accustomed to word (token) pairs to detail the incidents that are both concise and clear.
- The association rules of tweets require less computing time than that of traditional web media or reports which have longer contexts than those of tweets. According to our study, the required computing time for results in Table 6-2 takes about 17 min with computers (i7 3720QM, 32G RAM). Short text length of tweets facilitates the association rule mining.
6.3 Comparing with the classification by sLDA

6.3.1 sLDA Process

In this section, we exploit the supervised Latent Dirichlet allocation (sLDA) (Mochihashi, 2009), a topic modeling method. As compared to the SVMs, topic modeling assumes that a topic is a probability distribution over a group of words (tokens) which describe a semantic theme and the features of a document can be divided into several different topics instead of different words (tokens). Thus, sLDA is capable of reducing the dimensionality of the words. As compared most of the topic models including Latent Dirichlet allocation (LDA) which are unsupervised, sLDA can infer latent topics of the response on the basis of a manual label. The major differences between unsupervised and supervised topic models are the techniques to reduce dimensionality. The advantages of sLDA have been proved in several studies. However, the effectiveness of tweets is under question mainly in first, compared to the data sources like film reviews (Boyd-Graber and Resnik, 2010), image (Rasiwasia and Vasconcelos, 2013), etc., tweets have fewer words and may not generate reliable topics; second, unlike topics like Named Entity (Xu et al., 2009), sentiment (Lin et al., 2012), etc., traffic-related topics are comparably less general.

According to Mcauliffe and Blei (2008), each tweet post and label are processed from the following generative process:

1. Draw topic proportions $\theta|\alpha \sim \text{Dir}(\alpha)$;
2. For each word
   
   (a) Draw topic assignment $z_n|\theta \sim \text{Mult}(\theta)$;
   
   (b) Draw topic assignment $w_n|z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$;
3. Draw response variable $y|z_{1:N}, \eta, \sigma^2 \sim \text{Mult}(\eta^T \bar{z}, \sigma^2)$. 

150
Where $Dir(\alpha)$ is the Dirichlet distribution; $Mult(\theta)$ is the multinomial distribution; $z_n$ is the topic of the word $w_n$ (token); $\beta_{zn}$ is the multinomial distribution parameter for $z_n$; $\bar{z} = \frac{1}{N} \sum_{n=1}^{N} z_n$. We follow the generative process and E-M procedure (Mcauliffe and Blei, 2008) to infer the unknown parameters in the topic and word distributions. We implement 5-fold cross validation in the process of model training as in Section 3.

6.3.2 Comparisons of classification results

In sLDA, there are only two latent topics (label): accident-related and non-accident-related. For the tweet post, the words are stemmed and tokenized and stop-words are filtered as discussed in Section 3. The results are shown in Table 6-3 and Figure 6-5. The topic words are tokenized before classification as discussed in Section 3.2.

The first 30 topic tokens generated by sLDA are shown in Table 6-3. In the table, Topic 2 refers to the tokens that are accident-related while topic 1 is the reverse. By comparing with the results of individual tokens in Figure 6-2, the positively correlated tokens are bolded and blue-colored whereas the negatively correlated tokens are shaded and red-colored. Those tokens in topic 2 list contain most of the tokens that have a positive relationship with the accident label while some in topic 1 are those with a negative relationship. One can admit that to some extent, sLDA properly classify the tweets that are accident-related but the precision may be greatly influenced by the tokens that have no specific meanings.

Table 6-3 Lists of top 30 topic words (tokens) of sLDA models on tweets
Topic 1  like im damag incid accident dont shit can tailgat will u lol roll time peopl dai hi make cant hit wa onli love thing still control life call lost someon

Topic 2  car accid traffic lane got bad wtoptraff block befor road rd close todai va involv almost 3 66 polic 2 95 two near left okai bu major 1 front

Figure 6-5 compares the regression results between sLDA and SVMs with paired token features. In SVMs, support and confidence are set to be 0.01 and 0.6 respectively. The results of sLDA are slightly lower than SVMs. The precision is better than that in (Gu et al., 2016) mainly because there are 5 different categories (labels) in that study. However, the precision for accident-related is too low indicating that a number of tweets are predicted wrongly as accident-related by sLDA.

Figure 6-5 Comparisons of accuracy and precision between SVMs and sLDA

6.4 Improvements of classification by traffic-related information

In principle, the fusion of multi-source data provides significant advantages over single source data (Hall and Llinas, 1997) and the integrations of association features inherent in
the tweet contents and other data sources are expected to produce more synthetic and informative results. As the traffic accident potentially influence the road traffic operations, the abnormal patterns of traffic-related information are potential features. It is also a viable method of monitoring the traffic operations in traditional studies (Coifman et al., 1998; Oh et al., 2001). However, two major challenges exist in fusing the traffic data: first, the impact of a traffic accident to its surround areas is unknown both in time and geographic scale; second, the traffic flow patterns are difficult to identify given the large volumes of historical data. Here we employ a flow-occupancy pattern method to extract the traffic-related features.

6.4.1 Flow-occupancy pattern modeling

According to the thorough study of the fundamental diagram (Jin and Ran, 2009), it is widely accepted that there exists a relationship between the traffic flow and occupancy (or density). Previously studied in Zhang et al. (2016e), the recurrent flow-occupancy pattern of each detector can be unveiled by studying the historical traffic volume and occupancy data. For each detector, we evenly divide the traffic occupancy into $N$ separate groups. For each group of traffic occupancy, we define the median of the corresponding traffic flow values as the traffic signature, since it is less affected by outliers than mean. The traffic signature of a detector $d$ is defined as the vector of these traffic flow values. That is $\mathbf{F}^d = (F_1^d, F_2^d, \ldots, F_0^d, \ldots, F_N^d)$.

Where $F_0^d$ is the median value of traffic flow given a range of occupancy $o$ in detector $d$. One can see that for each detector, the flow-occupancy pattern is a vector of $N$ traffic flow values. If there is no traffic flow record over a certain occupancy, we employed the linear interpolation of traffic flow median of adjacent occupancies. We can finally obtain
the traffic signatures of more than 15,000 detectors in over 1,250 signalized intersections. In this chapter, $N$ is set as 50.

It is well known that the flow-occupancy relationship is diverse (e.g. triangle, parabola, trapezoid, broken-line, etc.). In our study, we do not make any assumptions about this relationship. Instead, we assume the nature of the relationship remains unchanged:

- **Assumption 1**: there exists an unchanged traffic signature in a given location. The traffic flow corresponding to a certain occupancy interval will mostly fall into a reasonable range, and those that deviate from the feasible range are regarded as traffic outliers.

To validate the assumption, we employ the K-means algorithm without pre-defining the clustering centers and the number of clusters to reveal the relationship. K-means clustering algorithm can partition the traffic signatures into finite groups of similar patterns. The inputs are the traffic signatures of all detectors, and the outputs are the collection of cluster centers and the cluster IDs that detectors belong to. We employ Akaike information criterion (AIC) (Akaike, 1998) to find the proper number of clusters. AIC measures the relative quality of the clustering results, shown in Equation (7).

$$AIC = \sum_{i}^{k} \sum_{d \in dom(i)} d(F_{di}, C^i) + k \cdot N$$

Where $F_{di}$ denotes the traffic signature of the $d$th detector that belongs to $i$th cluster. $C^i$ is clustering center of the $i$ th cluster. $d(F_{di}, C^i)$ is the Euclidean distance between traffic signature $F_{di}$ and its clustering center $C^i$. $dom(i)$ is the domain (collection) of all detectors in $i$ th cluster. $k$ is the current number of clusters. $N$ is the count of elements in a traffic signature, which equals to 50 in our study.
Our algorithm starts with the lower bound of the number of clusters and iterates the K-means clustering by increasing the cluster number. We calculate the AIC difference between the current iteration and the previous one. The iteration ends until the AIC difference is fewer than $\epsilon$. The algorithm is as follows:

Algorithm:

**Input:** The maximum number of clusters $K$, and traffic signature $F^d$ for all detectors. (in this chapter, our data has more than 15,000 rows and 50 columns. Each row represents the traffic signature of a detector.)

**Output:** Centers of clusters $(C^1, \ldots, C^i, \ldots, C^k)$;

Cluster IDs detectors belong to.

Assign the initial number of clusters $k=2$, initialize $\text{AIC} = +\infty$

Repeat

Implement K-means clustering algorithm with $k$ clusters:

Pick randomly the cluster centers $(C^1, \ldots, C^i, \ldots, C^k)$;

Repeat

Cluster each traffic signature $F$ to the nearest cluster center $C^i$ with $\min(d(F^d, C^i))$;

Replace $C^i$ by $\text{mean}(F^d)$;

Until none of the detectors switch clusters;

Calculate the AIC difference between each cycle;

Until AIC difference $\leq \epsilon$ or $k = \kappa$

The AIC values will theoretically decrease with the increase of $k$. In this chapter, we set $\epsilon$ as 3%. When $k=15$, the change in AIC goes lower than 3%, as shown in Figure 6-6.
Figure 6-6 AIC values for different number of clusters

Figure 6-7 15 different clustered centers of traffic signatures
Thus, we finally cluster nearly 15,000 detectors into 15 different groups. The centers of clusters are shown in Figure 6-7. From the shape of our clustering results, it is not surprising that the relationships between traffic flow and occupancy differ greatly from each other. Unlike a predefined relationship, this method has certain advantages:

- The method is totally driven by the analysis of large-scale data. The aggregation analysis of large-scale data can lead to reduced noise in the results.
- The method clusters the traffic signatures with similar flow-occupancy patterns and potentially identifies the location of detectors that hold similar characteristics in the road network.
- The method excludes the influences of daily differences or time-of-day differences inherited in the traffic data.

6.4.2 Abnormal pattern identification and traffic-related features

The output cluster centers represent the normal relationship between traffic flow and occupancy. One can intuitively figure out the possible traffic outliers by comparing the clustered center to the original data as shown in Figure 6-8.

Figure 6-8 Comparisons between clustered centers and the original traffic flow and occupancy data in two sample detectors
For each cluster, the traffic flows over a specified occupancy interval are distributed around their cluster centers. Further, the outliers can be quantified by a probabilistic method that measures its deviation degree. Previous empirical examinations show that the distributions of the traffic flow in a particular cluster and occupancy interval follows a Gaussian distribution (Zhang et al., 2016b). The traffic outliers can be intuitively identified in the distribution tail.

Thus for each detector, the abnormal degree of traffic-related data can be quantified by the cumulative probability of the distribution.

\[
P_{dt} = \Phi \left( \frac{F_{odt} - C_{oi}}{\sigma_{oi}} \right)
\]

Where \(P_{dt}\) is the probability for detector \(d\) over time period \(t\). \(i\) indicates the \(i\)th cluster of \(d\); \(F_{odt}\) is the traffic flow data over traffic occupancy interval \(o\); \(\sigma_{oi}\) and \(C_{oi}\) is the standard deviation and center of traffic flow in Cluster \(i\) over occupancy interval \(o\). \(P_{dt}\) quantifies abnormal probability for the deviation of traffic data from its cluster centers. Equation (8) captures the deviation of traffic flow to its clustering center, both decreased and increased traffic flow over certain time period \(F_{odt}\) can result in a higher probability value \(P_{dt}\). Thus, the larger \(P_{dt}\) is, the worse the traffic operations should be and the more likely the traffic is influenced by traffic accident. This probability can be employed as the traffic-related feature in our model.

In the process of extracting traffic-related features of a tweet, we mainly study the traffic related information within certain spatial (\(d\)) and temporal (\(t\)) ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. For each tweet, the corresponding abnormal probabilities will be further aggregated based on two major considerations: From
the geographic perspective, as the geographic impact of the traffic accident may vary, the increase of traffic probabilities may happen either in all places around the accident site or just only in partial places. From the temporal perspective, as the traffic accident may happen either before or after when the tweet is blogged, the increase of abnormal probabilities may happen either over the whole time period or just only a certain time span. According to these considerations, two traffic-related features are then generated for each tweet:

\[
p_{traffic} = \frac{1}{NUM} \sum_{t \in dom(t)} \sum_{d \in dom(d)} p_{dt}
\]  

\[
q_{traffic} = Q3(\{p_{dt}, d \in dom(d) \cap t \in dom(t)\})
\]

Where \( t \) is the hour period; \( d \) is the detector ID and \( i \) is the cluster ID; \( dom(d) \) is the domain of all the detectors within the geo-scale of the tweets and \( dom(j) \) is the domain of all time periods within the time-scale of the tweets; \( Q3() \) is the operator of 75th percentile; \( NUM \) is the total number of traffic observations related to a tweet. It is worth mentioning that both \( p_{traffic} \) and \( q_{traffic} \) are discretized before putting into the regression model.

6.4.3 Classification results with traffic-related information

We further conduct a comparison between the prediction results with and without traffic-related features. In SVMs, support and confidence are set to be 0.01 and 0.6 respectively. The results shown in Figure 6-9 indicate that traffic-related features do not necessarily improve the prediction results.
One possible explanation is that some traffic accidents may not generate impact on surrounding traffic, or the locations of the tweets do not match that of the accident sites. This needs to be further verified by comparing the accident-related tweets with the traffic accident logs.

6.5 Comparisons with the ground truth

6.5.1 Coverage of accident-related tweets

Even though the accident-related tweets in our study amounts to 422, it is still a very small number as compared to total 52,496 official accident records. It is worth mentioning that we only consider the geo-tagged tweets which take no more than 5% of all tweets posted online. One advantage of geo-tagged tweets to detect traffic accidents is that they can provide clear longitude and latitude information of where the tweets are posted. For those tweets without specific latitude and longitude messages, we can possibly infer their locations according to their tweet messages (Ikawa et al., 2012). However, it is obvious that not all non-geo-tagged tweets can provide enough location messages. Given the low
coverage of geo-tagged tweets and the disadvantages of non-geo-tagged tweets, one can see that tweets are still unlikely to cover all the traffic accidents with high probability.

Thus, the tweets are more probable to be a viable supplement rather than a replacement to the existing detection method. This is mainly because the they are relatively small-scaled incidents (Schulz et al., 2013) and seldom arouse public attentions. The influence of them may not be as high as that of earthquake or festival parades. And not all travelers are willing to leave a corresponding message online. Also, when passing by the site of a traffic accident, most of the drivers cannot tweet about it just for their own traffic safety.

6.5.2 Features of time and space differences between tweets and accidents

As the traffic-related information may not improve the prediction results, the traffic conditions around where the tweets are posted may not be significantly influenced. To find possible explanations, this chapter further examines the time and space differences between the accident-related tweets and corresponding accidents in the official log. Given an accident-related tweet, we extract the accident records from the log which is close to the tweet locations over a certain time window. The time window is the set 1 hour before and after the tweet time. The comparison results are insightful in studying the potentials of tweets in traffic accident detection.

According to our examination, of more than 400 labeled accident tweets, there are about 300 of them can be traced to an accident record by VDOT. It is also possible that one tweet corresponds to more than one accident records and there is no additional information for us to specify which one the tweet indicates. In this case, we choose the accident record that is the nearest to the tweet location. The time when the tweets are posted can be either earlier or later than the starting time of the traffic accident records as shown in Figure 6-10.
(a). Comparing with the starting time in the traffic accident log, nearly one-third of the accident-related tweets are posted earlier than the traffic accident. This coincides with the findings in (D'Andrea et al., 2015) that tweets detect traffic accidents much earlier than traditional media. If so, detecting the online accident-related tweets significantly reduces the response time.

Figure 6-10 Distributions of (a) time and (b) space difference between the accident-related tweets and the accident records by VDOT.

As shown in Figure 6-10 (b), the shortcomings are also obvious that the space differences are sometimes too large and it is hard for traffic operators to pinpoint the accident site solely with the latitude and longitude information of geo-tagged tweets. This also explains the reasons why traffic-related information cannot improve the prediction accuracy because the travelers sometimes tweet where they are far away from the accident sites (e.g. at a coffee shop). This increases the difficulties in real-time detection of traffic
accidents. The location information of an accident can sometimes be extracted from what people are tweeting (Gu et al., 2016) but the method may not work for all tweets.

6.5.3 Features of unrecorded tweet accidents

The comparisons between tweets and accident records also reveal that nearly a quarter of accident-related tweets express explicit meanings about the traffic accidents but cannot be traced by official accident logs. After our examinations, more than one-third of untraced tweets are from the media channels such as “wtop”, “wtoptraffic”, etc. The locations for these tweets may not provide useful locations for accident detection. Other tweets may possibly be accounted by several other reasons.

First, compared with the traffic accident log, it is entirely possible that the tweets can capture the unexpected small events happened in our daily life. These events may include those “mild” accidents that do not incur the attention of traffic police and thus may not be included in the official log. The consequences of these events such as the road lanes blocking or cars slowing down may not last long and the corresponding affairs may come with a proper handling. If so, the unrecorded tweets may act as a supplement of the current accident detection system. The examples of such tweets can include:

- “woooo got rear ended on i495 going to md great way to start a monday morning”
- “holy shit i just crashed my dads car”

Second, other reasons could include: 1) some of the accident-related tweets may be posted too far away from the accident site; 2) some tweet users retweet about an accident instead of seeing in person; 3) some tweeters may misjudge the situations and their
inferences are from the jammed conditions of the roadways. In sum, some tweets may just be false alarms that need further verifications. For example:

- “sooo the car just said attention there is a car accident 12 miles ahead wtf kin of car does that”
- “major vehicle accident southbound i95 near lorton va traffic dmv”

After comparing with the ground truth, it can be concluded that the tweets labeled by our model can possibly identify the existence of potential traffic accidents. Such identifications are typically faster than the traditional media. The locations of the traffic accidents may not be just exactly the latitude and longitude where the tweets are posted and the traffic operators should incorporate more information sources pinpoint the locations. In sum, it is entirely possible to increase the efficiency of traffic accident detection by monitoring the real-time social media data.

**6.6 Conclusions and discussions**

In this chapter, we employ the SVMs to detect the traffic accident from tweets. The prediction employs three important features: single token, paired token and traffic-related data to achieve a more accurate and effective on-site traffic accident detection. Also, the prediction results are compared with those of sLDA, a supervised topic modeling method, and the accident-related tweets are compared with the traffic management logs. Our major findings can be summarized as follows:

First, we thoroughly investigate the tweet contents related to traffic accidents. We found token features: single tokens and paired tokens that may indicate the event of traffic accident. Our results show that paired tokens can capture the association rules inherent in
the accident-related tweets and efficiently increase the accuracy of the traffic accident detection.

Second, we unveil the relationships between traffic flow and occupancy based on the fundamental diagram using a large-scale dataset and point out that these relationships vary across different locations. We employ the K-means clustering algorithm to cluster the detectors into different flow-occupancy patterns. The derived traffic-related information barely improves the results of accident prediction.

Third, the comparisons between the accident-related tweets and the traffic accident log provide the following insights: First, sometimes the tweet reports the traffic accident much faster than the traditional media, so detecting the online accident-related tweets can significantly reduce the response time. Second, tweets can capture those “mild” accidents that do not incur the attention of traffic police and this indicates the possibility of tweets detecting unreported accidents. Third, tweets can provide false alarms. Some tweets cannot give an accurate location of the accident site and more data sources should be further examined.

Finally, it is concluded that integrating social media data into the traffic-related study opens up a wide range of possibilities for research in crowd-sourced transportation. The results show that social media data might be noisy and even unreliable, so solely relying on social media data is still not a perfect option. Further studies can focus on the data fusion of different data sources to better realize the purposes of other research such as traffic jam detection, traffic emergency evacuation (Asamoah, 2014), passenger flow prediction (Ni et al., 2014), etc. The spatial-temporal features of traffic data are also worth studying for
regional traffic operations. Note that our tweet data are geotagged. It would be an interesting extension to detect traffic events with non-geotagged tweets.
CHAPTER 7 AN EXPLORATORY STUDY ON THE CORRELATION BETWEEN TWITTER CONCENTRATION AND TRAFFIC SURGE

Within NOVA area, we mainly investigate the tweets in July. Our incentives are intrigued by some preliminary examinations of the tweets. For instance, in February 2014, there was a keyword surge of “capitalweather”. People tweeted the delay caused by the “Biggest snow storm since Snowmageddon”. Public service info feed “Metrobus Info” lively broadcasted the congestion at North Capitol & New York Ave in DC area. This was widely retweeted by other users. One can say that the public events reflected by Twitter concentrations potentially exert pressure on the road network and aggravate the traffic congestion. To ensure that the tweets are collected from the general public, Twitter users from traffic authorities such as “I95VA”, “MKA_NVA” have been removed from our study. Also, some of the tweets from media or press such as “nbcwashington” etc. were also excluded after our empirical examination.

7.1 Twitter concentration study method

7.1.1 Twitter concentration extraction and filtering

We can automatically extract Twitter concentrations that have the following features:

- They are related to real events that many people witness and are willing to share their observations or experience via Twitter.
- They potentially have connections with on-road traffic-related activities and may involve some kind of traffic movement.
First, we extract the Twitter concentrations from the set of tweets from January 2014 to December 2014. In most of the cases, keywords contained in the tweet can differentiate it from other tweets. Our algorithm first splits the tweet texts into separate word characters that form a large word database. In the database, we first search for the keywords that frequently appear in each day. Then, for all the frequent words, we select the words that frequently appear in one day but not so frequently or even vanish in other days.

**Algorithm 2: Twitter concentration extraction**

**Input:** Tweets collections throughout 2014  
**Output:** Keywords of each day and tweets that contain the keywords

For each day  
- **Decompose** the tweets into vectors of words on that day  
- **Count** frequency $f_{w(d)}$ of each word $w$ on day $d$;  
- **Pick** the words that satisfy $f_{w(d)} \geq \epsilon \sum_w f_{w(d)}$  
- **For** all selected $w(d)$  
  - **Select** the words $w(d)$ as the keywords $k(d)$ that satisfy  
    \[
    \text{count}[w(d) \cap \text{dom}(w(d))] \leq \tau
    \]
  - **For** all $k(d)$  
    - **Extract** the tweets that contain $k(d)$ in July

We set the frequency threshold for the keywords by a ratio parameter $\epsilon$. $\text{count}[w(d) \cap \text{dom}(w(d))]$ counts the frequency of $w(d)$ in the domain of all $w(d)$, i.e. $\text{dom}(w(d))$ and whose frequency is no larger than $\tau$ is selected. For the selection of $\epsilon$, if we increase the value of $\epsilon$, it may miss some important keywords due to the large datasets. If we lower $\epsilon$, it may incur more computations. Our experiences show that there are not so much difference between 10% and values lower than 10%. The value of $\tau$ should not be
too large because longer periods of events will diminish the enthusiasm of the people and these events may not be a reflection of Twitter concentration. One may increase $\tau$ if their data covers more than one year because there may exist yearly events. In our study, we set $\epsilon$ and $\tau$ to be 10% and 3 respectively. By comparing the frequent words in different days, the stop-words such as “is”, “and”, “us”, etc. can be eliminated and the remaining frequent words are the keywords that may indicate a kind of social activity. Table 7-1 shows some keywords of the day in July and some possible related social events.

### Table 7-1 6 keywords and related social events in July 2014

<table>
<thead>
<tr>
<th>Date</th>
<th>Keyword</th>
<th>Social events</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/1/2014</td>
<td>waffles</td>
<td>Waffle House restaurant just tweeted the most American waffle breakfast possible</td>
</tr>
<tr>
<td>7/3/2014</td>
<td>louis</td>
<td>Louis Zamperini, an American war survivor in World War II, died</td>
</tr>
<tr>
<td>7/4/2014</td>
<td>freedom</td>
<td>Celebration activity such as Freedom Fest 2014 fireworks</td>
</tr>
<tr>
<td>7/18/2014</td>
<td>fotosdeprinceroyce</td>
<td>Prince Royce concerts</td>
</tr>
<tr>
<td>7/16/2014</td>
<td>wjmc2014</td>
<td>Washington Journalism and Media Conference</td>
</tr>
<tr>
<td>7/26/2014</td>
<td>silver</td>
<td>A new metro line: silverline, is opened</td>
</tr>
</tbody>
</table>

### Table 7-2 Transportation lexicon

<table>
<thead>
<tr>
<th>Accidents</th>
<th>Carpooling</th>
<th>Drive</th>
<th>Junc</th>
<th>Passenger</th>
<th>Seatbelts</th>
<th>Trains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival</td>
<td>Carriage</td>
<td>Driver</td>
<td>Junction</td>
<td>Passengers</td>
<td>Shuttle</td>
<td>Transit</td>
</tr>
<tr>
<td>Arrivals</td>
<td>Cars</td>
<td>Drivers</td>
<td>Junctions</td>
<td>Passing</td>
<td>Sidewalks</td>
<td>Transport</td>
</tr>
<tr>
<td>Arrive</td>
<td>Collision</td>
<td>Drives</td>
<td>Kilometer</td>
<td>Pedestrian</td>
<td>Speed</td>
<td>Transports</td>
</tr>
<tr>
<td>Arriving</td>
<td>Combustion</td>
<td>Driving</td>
<td>Kilometers</td>
<td>Pedestrians</td>
<td>Speeding</td>
<td>Trasport</td>
</tr>
<tr>
<td>Auto</td>
<td>Commute</td>
<td>Drop</td>
<td>Lane</td>
<td>Periods</td>
<td>Speedlimits</td>
<td>Travel</td>
</tr>
<tr>
<td>Automobile</td>
<td>Commuter</td>
<td>Eastbound</td>
<td>Licence</td>
<td>Petrol</td>
<td>Speeds</td>
<td>Travelcar</td>
</tr>
<tr>
<td>Automobiles</td>
<td>Commuters</td>
<td>Eastern</td>
<td>Line</td>
<td>Pickup</td>
<td>Standstill</td>
<td>Travelcard</td>
</tr>
<tr>
<td>Automotive</td>
<td>Commuting</td>
<td>Exhaust</td>
<td>Lines</td>
<td>Priced</td>
<td>Steer</td>
<td>Travelcards</td>
</tr>
<tr>
<td>Baggage</td>
<td>Congested</td>
<td>Exit</td>
<td>Link</td>
<td>Queues</td>
<td>Steering</td>
<td>Traveline</td>
</tr>
<tr>
<td>Bicycle</td>
<td>Congestion</td>
<td>Flows</td>
<td>Metering</td>
<td>Rd</td>
<td>Stops</td>
<td>Travelled</td>
</tr>
<tr>
<td>Bicycled</td>
<td>Connect</td>
<td>Freeflow</td>
<td>Motor</td>
<td>Ride</td>
<td>Taxi</td>
<td>Traveller</td>
</tr>
<tr>
<td>Bicycles</td>
<td>Connection</td>
<td>Freeway</td>
<td>Motorbike</td>
<td>Rider</td>
<td>Taxicabs</td>
<td>Travellers</td>
</tr>
</tbody>
</table>
Second, we filter the tweets extracted from the first algorithm to decide whether they are traffic-related. Compared with tweet classification, this algorithm is an unsupervised method that roughly estimates whether the tweets are traffic-related and works to shrink the size of the tweet collection. We hypothesize that the individuals describe their events by event-related words, and each traffic-related tweet should have one or more traffic-related words. A transportation lexicon is shown in Table 7-2 which is referred to (Gal-Tzur et al., 2014). We made some revisions by excluding some words that are related to air, water, railway traffic, etc. The tweets that contain at least one term in the lexicon are reserved otherwise discarded. We finally extracted 1179 candidate traffic-related tweets.

7.1.2 Twitter concentration classification and labeling

Twitter concentration classification is a supervised learning method that calculates the correlation between the tweets and traffic. We employed the logistic regression model,
which is first introduced in 1958 (Freedman, 2009), as our learning model to train and test the tweets. We first train the model and use the model to label the candidate tweets obtained in Section 5 is traffic-related. The model is as follows:

\[ F(X) = \frac{1}{1 + e^{\beta^T X + \beta_0}} \]

where \( X = (X_1, X_2, \ldots X_i \ldots X_m)^T \). \( X_j \) represents the vector of \( i \)th feature and there are \( m \) features in total. \( \beta = (\beta_1, \beta_2, \ldots \beta_i \ldots \beta_m) \) is the vector of coefficients of each feature. \( \beta_0 = (\beta_0 \ldots \beta_0) \) is the vector of intercepts. \( F(X) = (F(X_1), F(X_2), \ldots F(X_i) \ldots F(X_m)) \) is the vector of probability values of the dependent variable.

The classification method proceeds in the following steps:

First, we pick randomly 2000 tweets that contain one or more words in the Transportation Lexicon in Table 7-2 from the tweet collections of the whole year. We manually label them to judge whether they are traffic-related. The labeled results are taken as the ground truths as well as the dependent variables \( F(X) \).

Second, each tweet is further decomposed into separate word characters that are called “tokens” in our study. The tokens can be English character, number or even Latinized letters and are taken as the candidates of independent variables. There are more than 6000 tokens in total.

Third, we conduct a stop-word filtering on the candidate feature words. The stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). It can rule out the tokens that have no apparent linguistic meanings or significant event indications including articles, conjunctions, prepositions, pronouns, etc. The stop-word list we used referred to (Ranks-NL, 2015).
Fourth, we include those tokens that may correlate with the labels. The correlation benchmark we choose is phi coefficient (Cramér, 1999), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as $\phi$) between two variables $x$ and $y$ is calculated as:

$$
\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})}}
$$

When $x = 1$, $n_{11}$ and $n_{10}$ are the counts separately for $y = 1$ and $y = 2$; when $x = 2$, $n_{01}$ and $n_{00}$ are the counts separately for $y = 1$ and $y = 2$. Those tokens whose correlation coefficient $\phi$ are higher than 0.05 are selected. These tokens totaling 71 are taken as our covariate features $X$.

Fifth, we estimate the coefficients of the variables in the regression model by maximum likelihood estimation (MLE). This likelihood estimation can be realized by an iterative process such as Newton’s method (Ryaben’kii and Tsynkov, 2006) and the estimation of both coefficient values and significance are detailed in (Cohen et al., 2013). To increase the accuracy of the predicted model, we implement 5-fold cross validation (Geisser, 1993), which is a popular model validation method. Cross-validation can give insight on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data set. We finally obtained an overall estimation by averaging 5 test results. The accuracy of the model is 0.76.

Finally, the prediction model obtained in the previous step is employed to test the candidate traffic-related tweets obtained in Section 5. In our study, we take $F(X_i)$ as the
traffic accident probability of $i$th tweet data. The results show that of all 1179 tweets from the first classifier, 164 tweets may correlate with the traffic with $F(X_i) > 0.5$.

7.2 Traffic surge detection method

7.2.1 Traffic clustering algorithm

Besides the large-scale sensor data, the traffic data inheritably possess time-of-day features such as AM peak and PM peak. Thus, the traffic surge should be justified by a time-of-day clustering method. It is worth mentioning that we do clustering on the data separately on weekdays and weekends because the traffic conditions may be quite different. The following algorithm works almost the same for weekdays and weekends, and we do not intentionally distinguish that.

Backed by this setup, we start from detector level. The traffic occupancy data of detectors are collected every 15 minutes, and we take the median of the traffic data collections in different hour period as the traffic signatures of the detectors. That is:

$$O^d = (O^d_1, O^d_2, \ldots, O^d_j, \ldots, O^d_N)$$

where $O^d_j$ is the occupancy median in the $j$th hour period in the detector $d$. There are in total $N$ hour periods in one traffic signature $O^d$. As our study hour period is from 4:00 a.m. to 21:00 p.m., $N$ is set to be 18. The median value possibly eliminates the fluctuations of traffic data in different days and is less likely to be influenced by outliers than mean. Previous study argues that given the combination of direction, connectivity and locality of a road segment, one can distinctively determine the corresponding traffic signature (time-of-day features of speed) of a road segment with high probability (Banaei-Kashani et al., 2011). Enlightened by this idea, we also assume the following:
Assumption 1: there exist unchanged traffic signatures in a given detector. The time-of-day traffic occupancy over a certain hour period fall into a reasonable range, and those that are obviously higher from the feasible range are traffic surge.

The traffic signatures in more than 15000 traffic detectors constitute the raw database. To find the feasible occupancy range of each hour period, we employed the K-means algorithm with a principled way of finding the number of clusters and the cluster centers. This algorithm can partition the traffic signatures into finite groups of similar patterns and output the centers of clusters as well as the cluster IDs detectors belong to. The algorithm is shown as follows:

**Algorithm 1: Traffic signature clustering**

**Input:** The maximum number of clusters $\kappa$ and the traffic signature matrix $O^d$ (in this chapter, this matrix contains 15000 rows and 18 columns. Each row is the traffic signatures of detectors).

**Output:** Centers of clusters $(C^1, ..., C^i, ..., C^k)$;

The cluster IDs detectors belong to.

Assign the initial number of clusters $k=2$, initialize $AIC=+\infty$

Repeat

Implement K-means clustering algorithm with $k$ clusters:

Pick randomly the cluster centers $(C^1, ..., C^i, ..., C^k)$;

Repeat

Cluster each traffic signature $O^d$ to the nearest cluster center $C^i$ with $\min(d(O^d, C^i))$;

Replace $C^i$ by $mean(O^d)$;

Until none of the detectors switches clusters

Calculate the ratio of AIC change: $diff(AIC)/AIC$

Until $diff(AIC)/AIC \leq \epsilon$ or $k = \kappa$
In the algorithm, \( \epsilon \) is a threshold value set to be 3\% in this chapter. The Akaike information criterion (AIC) (Akaike, 1998) is employed to measure the relative quality of the clustering results.

\[
AIC = \sum_{i}^{k} \sum_{d \in \text{dom}(i)} d(O^{di}, C^i) + K \cdot N
\]

where \( O^{di} \) denotes the traffic signature of the \( d \)th detector that belongs to \( i \)th cluster. \( C^i \) is the center of the \( i \)th cluster. \( d(O^{di}, C^i) \) is the Euclidean distance between a traffic signature \( O^{di} \) and its cluster center \( C^i \). \( \text{dom}(i) \) is the domain (collection) of all detector ID whose traffic signature belongs to \( i \)th cluster. \( k \) is the current number of clusters. \( N \) is the count of elements in a traffic signature which equals to 18 in our study. Theoretically, the smaller the \( AIC \) is, the better the clustering result should be. For computational efficiency, the algorithm stops when the increase of cluster number brings no more than 3\% additional benefits. The AIC results are shown in Figure 7-1. When \( k=15 \), \( \text{diff}(AIC)/AIC \) goes lower than 3\%.

There recommend two important criteria in selecting the number of clusters: First, the cardinality of small size clusters may decrease with the increasing of the number of clusters. The cardinality should not be too small because corresponding results from clusters of large cardinality produce more reliable cluster centers in the later study. Second, we may also use BIC or other statistics to measure the quality of clustering. Same as that of AIC, the results of other statistics do not also indicate an unconstraint large number of clusters.
Figure 7-1 AIC values for different $k$

The clustered centers are shown in Figure 7-2. From the shape of our clustering results, it is not surprising that different cluster centers vary not only in shapes but also in scales. It reveals a clear time-of-day feature for each detector. This method can find the outliers in the traffic occupancy due to several of its advantages:

- The method fully considers the time-of-day features of traffic patterns inherited in the traffic data.
- The method is totally based on the field data which is in large scale. The aggregation of large-scale data may eliminate the possible noises from the results.
- The method clusters the traffic occupancy only in July, which can diminish the effects of traffic operations in different months.
7.2.2 Traffic surge definition in detector level

For each cluster, the traffic occupancy over a specified hour period should be distributed around their cluster center. An outlier is far away from the cluster center, and its level of deviation from the center can be justified by calculating its probability. We empirically check the distributions of traffic occupancy in all hour periods in different clusters, and two of them are shown in Figure 7-3. After reviewing the empirical
distributions, we can conclude they reasonably approximate them with a normal distribution.

Figure 7-3 Distribution of traffic occupancy in two clusters at 7:00 p.m.

Thus, we can justify the severity of traffic increase based on the normal distribution.

We define the traffic surge probability as probability \( \Phi \left( Z = \frac{O_j^d - C^i_j}{\sigma^i_j} \right) \) during the \( j \)th hour period in detector \( d \) and cluster \( i \). Here \( C^i_j \) is the cluster center in Cluster \( i \) during hour \( j \); \( \sigma^i_j \) is the standard deviation of traffic occupancy in cluster \( i \). The closer the value is to 1, the greater the traffic surge should be.

7.3 Correlations between Twitter concentrations and traffic operations

For each tweet, we mainly study the traffic related information within certain spatial and temporal ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. It is worth noting that:
• Public activities related to Twitter concentrations may happen either before or after when the tweet is blogged. So does the traffic surge.
• As the geographic impact of public activities may vary, the traffic surge may exist in one or even more intersections nearby.

Thus, influenced by public activities related to the Twitter concentrations, there are mainly two different traffic surges: traffic surge in part of the detectors or over shorter time periods; traffic surge in most detectors and over long time periods. The first kind of traffic surge can be justified by the 75th percentile value of traffic surge probability:

\[ q_{traffic} = Q3\left( \{ \phi \left( Z = \frac{O_{dj}^i - C_{ij}^l}{\sigma_j^l} \right), d \in \text{dom}(d) \cap j \in \text{dom}(j) \} \right) \]

Where \( j \) is the hour period, \( d \) is the detector ID and \( i \) is the cluster ID. \( \text{dom}(d) \) is the domain of all the detectors within the geo-scale of the tweets and \( \text{dom}(j) \) is the domain of all time periods within the time-scale of the tweets. \( Q3() \) is the operator of 75th percentile. As this kind of traffic surge is dramatic in only part of the detectors while relatively mild in other detectors, value of \( q_{traffic} \) corresponding to a tweet should be relatively high to justify a traffic surge.

The second kind of traffic surge can be justified by averaged traffic surge probability:

\[ p_{traffic} = \frac{1}{NUM} \sum_{j \in \text{dom}(j)} \sum_{d \in \text{dom}(d)} \phi \left( Z = \frac{O_{dj}^i - C_{ij}^l}{\sigma_j^l} \right) \]

Where \( NUM \) is the total number of traffic occupancy data related to a tweet.
For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by a threshold value of either $p_{traffic}$ or $q_{traffic}$. Here is our assumption on the detection of traffic surge from Twitter concentrations:

- **Assumption 2**: For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by either $q_{traffic} \geq q_{traffic}^0$ or $p_{traffic} \geq p_{traffic}^0$, where $q_{traffic}^0$ and $p_{traffic}^0$ are two parameters.

Given Assumption 2, the public events of Twitter concentrations can impact the surrounding traffic, and this impact can be justified by these two traffic surge probabilities. Two important findings are worth mentioning:

- Empirical results show that the impact of different threshold values of $q_{traffic}$ on the result is low and even negligible.
- If we set the threshold of $q_{traffic}$ to be 0.8, the percentage values of traffic-justified Twitter concentration events may change with $p_{traffic}$ as shown in Figure 7-4. Given $q_{traffic}^0=0.8$ and $p_{traffic}^0=0.5$, 127 out of 164 Twitter concentrations (77.4%) can be justified by traffic surge.
Figure 7-4 Percentage of traffic-justified Twitter concentrations under different threshold $p_{traffic}^0$

Different threshold values $p_{traffic}^0$ may influence the final results. It is obvious that a higher $p_{traffic}^0$, indicating a more serious traffic surge condition, may correspond to a public event that arouses more Twitter concentrations. It will be of great use in future study to further explore and quantify the severity levels of traffic surge using Twitter concentration.

Table 7-3 presents two Twitter concentrations and its corresponding public events. Figure 7-5 illustrates the time-of-day fluctuations of traffic surge probabilities of these two Twitter concentrations. One can see that the overall traffic surge probabilities in Figure 7-5 (a) and (b) are above 0.5 for Twitter concentration (1). As a comparison for (2), traffic surge probabilities in one detector are high (see Figure 7-5 (d)), but low in the other (see Figure 7-5 (c)). This figure characterizes the influence levels of different Twitter concentrations in different geographic scales. In Table 7-3, keywords “silverline” and “4thofjuly2014” can justify the correlation between tweets and major public events. The Twitter concentrations indicate the occurrence of traffic-related activities that result from the public events. The results prove the potentials of Twitter concentrations in detecting the traffic surge. One can see that without knowing the type of public events in advance, detecting the traffic-related Twitter concentrations assists in interpreting the causality of traffic surge and provides insights for better decision-making in urban traffic management.

Table 7-3 The keywords, Twitter concentrations and public events corresponding to Figure 7-5
<table>
<thead>
<tr>
<th>Keyword</th>
<th>Twitter concentration</th>
<th>Public events</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) silverline</td>
<td>waiting at Wiehle to ride silverline</td>
<td>Silverline metro opened on July 26</td>
</tr>
<tr>
<td>(2) 4thofjuly2014</td>
<td>waiting for my friend to get here so we can roll out easternshore, 4thofjuly2014</td>
<td>Celebration activities on Independence Day</td>
</tr>
</tbody>
</table>

Figure 7-5 The traffic surge probabilities at different time-of-day that are related to Twitter concentration (1) and (2), listed in Table 7-3. Figure (a) and (b) represent two detectors.
associated with Twitter concentration (1), Figure (c) and (d) another two detectors with Twitter concentration (2). The dash line indicates when the tweet is blogged.

7.4 Conclusions and discussions

Our study investigates the correlation between Twitter concentrations and the traffic surge on July 2014. The results prove the potentials of using tweets to detect the traffic surge within a given scale of space and time. First, the traffic occupancy over a certain period may follow a normal distribution, and this feature is fully exploited to derive the probability that quantifies the traffic surge. Second, the correlation between Twitter concentrations and traffic surge indicate that the major social activities that are related to traffic could possibly deteriorate the nearby traffic congestions. Our experiments show that 77.4% of traffic-related Tweeter concentrations can be justified by local traffic surge.

Following these findings, one may further the study by tackling the limitations of the current approach. In our study, the tweets are collected through Twitter Streaming API with geo-location filter and cannot possibly cover all the traffic surge of the whole region. This may be due to the limited volume of geo-tagged tweets. Also, the classification method employed in our study may be limited by the size of the training datasets, and the precision of the results may increase by incorporating more tweets.

The potential applications of our study are also promising: First, the traffic surge detection algorithm is built on the “big data” analysis of previous data collections. It can precisely unveil the traffic patterns in a large road networks and even identify the anomaly traffic conditions. Second, this chapter can help traffic operators understand the cause of traffic surge and improve short-term prediction of traffic congestion (especially non-recurrent congestion) on roadways in the future. Third, the Twitter concentration can
broadcast the traffic-related events in a much more timely and quickly manner than traditional broadcasting media. Monitoring the social media data may deliver useful traffic event information, including traffic accident, traffic jam, road construction, etc. It will also be interesting to analyze the spatial-temporal correlations between traffic patterns and Twitter concentrations in future research.
CHAPTER 8 AN EMPIRICAL STUDY OF TRIP PURPOSE INFERENCE WITH CONNECTED VEHICLES TRAJECTORIES, LAND USE DATA, AND SOCIAL MEDIA DATA

8.1 Data preprocessing and empirical findings

8.1.1 Trip information from tweets

Our studies are enlightened by 5-month tweet collections in Ann Arbor from Jan. 1 to May 31 2016. There are more than 80,000 geo-tagged tweets and each entry is comprised of (1) User ID and Username, (2) tweet contents, (3) posting time, (4) latitude and longitude together with (5) the name of the Census-designated places (CDP). The tweet contents are the reflection of what people are interested in at the specific time and location and some of them express the clear indications of trip-related information as shown in Table 8-1.

Table 8-1 Sample tweets with clear trip-related information

<table>
<thead>
<tr>
<th>Tweet content</th>
<th>Location name</th>
<th>Google Place type</th>
</tr>
</thead>
<tbody>
<tr>
<td>im at cafe felix  wix in ann arbor mi</td>
<td>cafe felix  wix</td>
<td>Bistro</td>
</tr>
<tr>
<td>had the best time at paint and pour with these ladies</td>
<td>paint and pour</td>
<td>Art School</td>
</tr>
<tr>
<td>im at alley bar  a2alleybar in ann arbor mi</td>
<td>alley bar</td>
<td>Cocktail Bar</td>
</tr>
<tr>
<td>i got a little over excited at 5 apps  pacific rim</td>
<td>pacific rim</td>
<td>Asian Restaurant</td>
</tr>
<tr>
<td>drinking a mango ipa by beer grotto at beergrotto</td>
<td>beer grotto</td>
<td>Bar</td>
</tr>
<tr>
<td>im at black pearl in ann arbor mi</td>
<td>black pearl</td>
<td>Seafood Restaurant</td>
</tr>
<tr>
<td>im at vaultofmidnight in ann arbor mi</td>
<td>Vault of midnight</td>
<td>Book Store</td>
</tr>
</tbody>
</table>

Our empirical examinations and basic statistics analysis open up the possibility of employing the tweets to labeling the trip purposes:
First, the population of users in our tweet database amounts to 5,000 which is large enough to represent the heterogeneity of interests in the Ann Arbor region. At the same time, the geographic coverage of the tweets can be as large as the City of Ann Arbor and its surrounding places. One can even find the time and location distributions of the local people when posting tweets as shown in Figure 8-1 (a) and (b).

Figure 8-1 (a) The frequencies of tweet posting time and (b) the frequencies of tweet locations from Jan.1 to May.31 2016 throughout the geography of Ann Arbor region; (c) Month distribution of the same tweet contents from different users.
Second, the Twitter check-in data from different travelers may share the same format. This function can be jointly realized by “Foursquare”, etc. which ties the real-time location specific information by rewarding users for “check-in” to real sites at any location worldwide (Kietzmann et al., 2011). Figure 8-1 (c) shows that in different months can find the exact same tweets indicating the same location, but from different tweet users.

Third, the tweets are representatives of social awareness towards the public activities and social communications. One can assume that the frequencies of tweet locations as shown in Table 8-1 may, to some degree, represent the people’s location preferences. This is true when the tweet data are from thousands of diverse and distinct users.

Above all, one can see the text structures of a tweet when it gives a “location report” and thus the great potentials of tweets in detailing the land use information of locations. Even though, they may not convincingly deliver clear indications of trip origins, destinations and trip trajectories as people may tweet at any other places during the trips. Thus, in similar studies which use data sources other than tweets: such as bank notes (Brockmann et al., 2006), cellphone tower records (Song et al., 2010), the researchers tend to name the distance between two time-consecutive locations as “displacement” rather than “trip”.

8.1.2 Trip information from Connected-Vehicle

At the same time, we introduce into our study the Safety Pilot Model Deployment (SPMD) data which is acquired from the U.S. Department of Transportation’s (USDOT) Real-Time Data Capture and Management program. The trip trajectory information of the Connected Vehicles are recorded by the onboard devices which are primarily equipped to test the vehicle-to-vehicle communications and other automatic safety functions. There are
more than 300 drivers involved. After data preprocessing, we extract nearly 240,000 trips and the time distribution of these trips are similar to that of tweets as shown in Figure 8-2 (a) which is similar to that of tweets as shown in Figure 8-1 (a). The recorded GPS locations reveal the trajectories of the vehicles as shown in Figure8-2 (b).

Figure 8-2 (a) The frequencies of all trip arrival time (b) Trip trajectories of two Connected Vehicles; (c) All trip ends.
As the data is recorded in a millisecond resolution and is in a large amount, Figure 8-2 (b) does equal samplings on the data for display purposes to reflect the velocity. The scarcity of points indicates high operating speed while density means just the reverse. We further extract the trip ends where the engine shut down. Similar to tweets, the geographic distribution of trip ends shown in Figure 8-2 (c) in the Connected Vehicles dataset have even more distinct geographic features. Compared with the tweet locations shown in Figure 8-1 (b), the trip ends can be covered by the locations extracted from tweets. As the trips usually end on the roadside parking or in the middle of a plaza, it is difficult to interpret underlying trip purpose if there are more than one places nearby.

8.1.3 Problem statement and study flow

In this chapter, we infer detailed trip purpose information of Connected Vehicles by combining the trip trajectories and the tweet locations. For proper classification purposes, we even introduce the land use data from the public GIS data sources of “City of Ann Arbor” (2015) and Google Place Types of the facilities. Our research process can be illustrated as in Figure 8-3:
Figure 8-3 Flow chart of our study.

8.2 Trip purpose inference

8.2.1 Tweet location extraction

There are two major issues to be addressed before we can associate the tweet location information to trip information:

- How to automatically recognize the semantic indications of a place?
- How credible are the tweet GPS locations and its corresponding location information?

Location inference from microblog messages has been intensively studied in (Cheng et al., 2010; Gu et al., 2012), etc. Through a supervised learning method or other language processing algorithm, most studies can reach good precision results by training and learning properly preprocessed datasets. Despite language differences, researchers have found typical sentence structures in place-related tweets. According to the features of Twitter location services, tweet users check-in using a uniformed message format. For example, the “Foursquare” location service uses the format of “I’m at (location address)” (Ikawa et al., 2012). Furthermore, we find that people are accustomed to using certain prepositions such as “at”, “on”, “in”, “by”, etc. when check-in tweeting.

Thus, we first collect all tweets around the trip ends of the Connected Vehicles. For all these tweets, we screen out those containing the selected prepositions and then extract from those tweets the first four words following those prepositions. In this chapter, we only care about the tweets around the trip ends as they are more likely to have travel-related topics. Also, the following manual examination on these tweets can be more effective and quickly as most irrelevant tweets are filtered out.
Through a web search, four words following the prepositions are enough to indicate, if any, a place. We made a verification through manual examination to see whether they have clear place indications. The results in Table 8-2 (a) show that in a tweet content, examining the words following preposition “at” is the easiest way to find a location information. Explanations for this may be various including the language customs, semantic indications, features related to the city of Ann Arbor, etc.

We further examine the distance between the places’ “tweet locations” and the real geographic locations which are obtained by online searching. The small distance between the two locations indicates that the tweet users are reporting at or near the real place locations, and thus his tweet can be used to label the trip purposes while the long distance just indicates a false alarm. From Table 8-2 (b), one can see that taking into consideration of the GPS errors and the occupied area of the places, some tweet check-in data is accurate enough to give a good location indication and most of them are credible. One can even see that “at” has the relatively small distances. The distance checking proves the credibility and accuracy of some tweets in reporting a place. Setting the distance threshold to be 200 meters, one can obtain nearly 230 qualified tweet locations. For those facilities, we record the corresponding real coordinates for further analysis.
Table 8-2 (a) Statistics of tweets around the trip ends and those with clear indications of locations and (b) Quartile values of distance between the tweet locations and the real geographic locations

(a)

<table>
<thead>
<tr>
<th></th>
<th>above</th>
<th>at</th>
<th>behind</th>
<th>below</th>
<th>besides</th>
<th>between</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count1*</td>
<td>12</td>
<td>477</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>707</td>
</tr>
<tr>
<td>Count2*</td>
<td>0</td>
<td>298</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Count3*</td>
<td>0</td>
<td>218</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>inside</th>
<th>on</th>
<th>opposite</th>
<th>over</th>
<th>to</th>
<th>under</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count1*</td>
<td>0</td>
<td>256</td>
<td>0</td>
<td>25</td>
<td>414</td>
<td>10</td>
</tr>
<tr>
<td>Count2*</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Count3*</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Count1*: Number of tweets containing the preposition.

Count2*: Number of tweets containing the preposition that have a location indication.

Count3*: Number of tweets containing the preposition with distance between the tweet locations and the real geographic locations less than 200 m.

(b)

<table>
<thead>
<tr>
<th></th>
<th>Q1 (m)</th>
<th>Q2 (m)</th>
<th>Q3 (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>17.34465</td>
<td>37.18043</td>
<td>247.2212</td>
</tr>
<tr>
<td>in</td>
<td>109.5387</td>
<td>190.5361</td>
<td>3053.727</td>
</tr>
<tr>
<td>on</td>
<td>10.47668</td>
<td>23.60871</td>
<td>516.8302</td>
</tr>
<tr>
<td>to</td>
<td>31.91862</td>
<td>138.03946</td>
<td>2260.963</td>
</tr>
</tbody>
</table>

Besides the real coordinates, by Google Places API, we also obtain the Google Place Types which is assigned to the facilities. We match the Google Place Type with our tweet
locations and thus each location has a category. According to our examinations, there are more than 100 different place types and the most frequent 5 place types are “Restaurant” (9.65%), “American Restaurant” (4.39%), “Bar” (3.51%), “Farmers' Market” (2.19%) and “Business School” (2.19%). Noting that “Restaurant” category is different from “American Restaurant” or others, crowdsourcing data of tweets and Google Place Types can deliver more detailed trip information.

8.2.2 Trip purpose inference with tweet locations

Within a certain space range, a trip end may have more than one surrounding tweet locations. Table 8-3 (a) shows the basic statistics of tweet locations and trip after we pair those two according to different space range settings. The percentage of trip ends with tweet data is low (2-3%) because the test bed, City of Ann Arbor, is a small town. The situation can be greatly improved in a big metropolitan area such as that of New York City or Washington D.C. The average number of tweet locations per trip is high enough to give a good trip purpose indication. The most frequently visited tweet locations as shown in Table 8-3 (b) also reflect the overall trip interests in City of Ann Arbor.

Table 8-3 (a) Statistics of trip destinations with tweet locations; (b) The most frequently visited tweet locations around trip ends

(a)

<table>
<thead>
<tr>
<th></th>
<th>Space range setting (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Number of trip ends that can be labeled</td>
<td>4896</td>
</tr>
<tr>
<td>Percentage of trip ends with trip locations</td>
<td>2.08 %</td>
</tr>
<tr>
<td>Average number of tweet locations per trip end</td>
<td>6.3</td>
</tr>
<tr>
<td>Average distance between facilities and trip ends</td>
<td>33.6</td>
</tr>
</tbody>
</table>
Based on the trip and location information available, we mainly study four important features of the surrounding tweet locations for an estimation of the trip purpose. These include: for a trip end,

- The visiting times of each tweet location category $Freq$: higher visiting frequencies in the past records indicates a higher public preferences of the locations.
- The average distance from a tweet location from that of the trip $Dist$ (meters): nearer locations are more likely to be the true trip end as people are more likely to park near where they want to go.
- The average time lags of a tweet location from that of the trip $Time$ (second): tweets with smaller time lags can be more representative.
- The average time-of-day difference of a tweet location from that of the trip $TOD$ (hour): tweet locations are more representative if they are posted in the same time-of-day.

Each feature of tweet locations is ranked from the best to worst, and a weighted average score is calculated as shown in Table 8-4:

$$Score = a_1 R(Freq) + a_2 R(Dist) + a_3 R(Time) + a_4 R(TOD)$$
Without further investigation, we can scientifically determine the values of $a_i$ to make an estimation. In this chapter, we set equal weights for all features. We take both location categories as trip purpose locations if there are equal scores. One can see that the trip purpose of Table 8-4 is more likely to be the “Association or Organization” than other locations. For each trip, we do the same ranking and finally obtain the corresponding trip purpose.

Table 8-4 Feature ranks of tweet locations around a specific trip location

<table>
<thead>
<tr>
<th>Tweet location</th>
<th>Features of tweet locations</th>
<th>Feature ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>Dist (m)</td>
</tr>
<tr>
<td>Association or Organization</td>
<td>3</td>
<td>169.2</td>
</tr>
<tr>
<td>Eclectic Restaurant</td>
<td>4</td>
<td>200.0</td>
</tr>
<tr>
<td>Pet Store</td>
<td>1</td>
<td>124.6</td>
</tr>
<tr>
<td>Restaurant</td>
<td>1</td>
<td>87.9</td>
</tr>
<tr>
<td>Tea House</td>
<td>2</td>
<td>162.2</td>
</tr>
<tr>
<td>Tea Store</td>
<td>1</td>
<td>121.3</td>
</tr>
</tbody>
</table>
8.2.3 Comparisons with land use data

In this subsection, we compare the tweet location categories and the land use categories. Figure 8-4 shows that the land use information is ambiguous and even incomplete as the categories such as “mixed use”, “commercial” and “office”, which take a large portion, do not distinguish from each other and may apply to the same facilities; also, “office” and “transportation” does not give much information about the trip purpose. Table 8-5 shows the trip purpose results for 7894 trips by both categories. Duplicate pairs of land use and tweet location categories are deleted, and the table gives a clear comparison.
Table 8-5 Comparisons of trip end labels between tweet categories and the previous land use categories

<table>
<thead>
<tr>
<th>Land use</th>
<th>Tweet location category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed Use</td>
<td>Cafe, Barbecue Restaurant, Video Arcade, Pizza Restaurant, Book Store, American Restaurant, Business School, Beer Store, Night Club, Performing Arts Theater, Graphic Designer, Bar, Transit Stop, Florist, Irish Pub, Brewpub, Library, Pizza Delivery, Spanish Restaurant, Beauty Salon, Noodle Shop, Hamburger Restaurant, Art Center, Antique Store, Cuban Restaurant, Restaurant, Chicken Wings Restaurant, Association or Organization, Asian Restaurant, Mexican Restaurant, Tea Store, Vintage Clothing Store, Tea House, Cocktail Bar, Ice Cream Shop, Art School, Mediterranean Restaurant, Clothing Store, Dentist, Eclectic Restaurant, Pet Store, 3-star hotel, Juice Shop</td>
</tr>
<tr>
<td>Public/Quasi-Public/Institutional/Organizational</td>
<td>Library, Brewpub, Chinese Restaurant, Chicken Wings Restaurant, Performing Arts Theater, Observatory, Barbecue Restaurant, Sports Complex, Vintage Clothing Store, Gym, 3-star hotel, Restaurant, American Restaurant, Public University, Art Center, Bed &amp; Breakfast, Bus Station, Tea Store, University, Cuban Restaurant, Antique Store, Irish Pub</td>
</tr>
<tr>
<td>Transportation/Communication/Utilities</td>
<td>Observatory, Tea Store, American Restaurant, Women's Clothing Store, Barbecue Restaurant, Brewpub, Sports Complex, Restaurant, Mexican Restaurant, Art Center, Library, Pet Store, Transit Stop, Grocery Store, Farmers' Market, Produce Market, Hamburger Restaurant, Chinese Restaurant, Chicken Wings Restaurant, Deli, Bus Station, Book Store, Performing Arts Theater</td>
</tr>
<tr>
<td>Office</td>
<td>Book Store, Barbecue Restaurant, Hotel, Brewpub, Pizza Restaurant, Association or Organization, Beauty Salon, Transit Stop, Art School, Chicken Wings Restaurant, Vintage Clothing Store, Restaurant, Performing Arts Theater, Mexican Restaurant, Hypnotherapy Service, Cocktail Bar, Spanish Restaurant, Tea Store, Bar, Irish Pub, Gym</td>
</tr>
<tr>
<td>Commercial</td>
<td>American Restaurant, Pizza Restaurant, Irish Pub, Book Store, Pizza Delivery, Barbecue Restaurant, Bar, Restaurant, Hamburger Restaurant, Graphic Designer, Association or Organization, Florist, Farmers' Market, Pub, Gym, Asian Restaurant, Antique Store, Bistro, Cuban Restaurant, Cocktail Bar, Art School, Brewpub, Pet Store, Night Club, Tea Store, Church, Hair Salon, Bagel Shop, Mediterranean Restaurant, Produce Market, Deli, Tuscan Restaurant, Food Products Supplier, Spanish Restaurant, Cafe, Beauty Salon, Beer Store, Video Arcade, Performing Arts Theater, Art Center, Italian Restaurant, Dentist, Noodle Shop, Chicken Wings Restaurant, Real Estate Developer, Business School</td>
</tr>
</tbody>
</table>

197
This comparison shows that:

- The land use data is sometimes ambiguous and overlap with each other. As compared, tweet location categories are with higher resolution and more informative.
- The tweet location categories also reflect public social preferences and popularity of activity locations as more frequent visited locations are valued as shown in Table 8-4.
- Seldom do people tweet about home. Table 8-5 depicts that all tweet locations are about commercial or non-profit facilities other than home.

More detailed place information enrich our understanding of the trip purposes: previous studies link land use codes and corresponding trip purposes by household survey. For instance: in (Wolf et al., 2001), land use category “Restaurant” links the trip purpose “eat”. In comparison, tweet location category can differentiate “Asian Restaurant”, “Seafood Restaurant”, etc.

8.3 Conclusions and discussions

In this chapter, an empirical study is conducted, as the first attempt, to infer the trip purpose by joining Connected Vehicles trajectories, land use data and social media data.
(Twitter). Our studies show that the tweets can report place names which can be extracted by first searching for the proposition “at”, “in”, “on”, etc. and then manual examinations. The analysis finds nearly 230 tweet locations that are verified by the real places within less than 200 meters. Moreover, those tweet locations are further categorized into 100 different Google Place Types. Also, the tweets can reflect the social preferences and popularities of an activity location as people may tweet about the places at different frequencies and periods. The tweet location category obtained from Google Place Types provide informative results for trip purpose inference study. They are shown to be more advantageous as compared with the land use data.

Three possible improvements are worth studying in the future: First, the timeliness of results need to be improved as the tweets were collected 2 years later than the trip data; and in the future, one can conduct human-involved experiments based on the first-handed participants’ trip records. Second, the features of the tweet locations should be further studied, and a supervised learning method can be employed for a more accurate prediction. This problem can be overcome by collecting the tweets and trip information from the same traveler. Third, the tweets can seldom identify the home which is a shortcoming in labeling the trip purpose. This can be solved by studying the corresponding trip trajectories and land use information as in (Bohte and Maat, 2009). Fourth, the studies can be more meaningful and accurate by incorporating the ground truth of the trip purpose. Only by this data source can we infer the trip purpose through a regression model and find influential factors. Overall all, with the increasing coverage of tweets, it can be expected that the result can be more accurate and representative in the future works. One can further explore potentials of
the tweets in trip purpose inference such as the relevance of tweet purpose category as to the real trip information, the trip pattern similarities between travelers and tweet users, etc.
9.1 Conclusions

This dissertation details the process of employing the traffic data, social media data, land use data and other data sources to deliver the advanced traveler information. The information includes:

(1) In Chapter 3, we propose to employ the **compression-based anomaly detection method** to effectively interpret the large collections of multi-dimensional traffic data.

We unveil the spatial and temporal traffic patterns in NOVA. The physical meanings of the anomalies and comparisons with other methods are validated. We first leverage dictionary-based compression theory for regional traffic flow pattern identification and derive an anomaly index to quantify the traffic patterns. By the method, two or more traffic-related metrics are interpreted in the same group, which prevents the complex definition of several different variables and parameters. The features of traffic flow patterns can be quantified as a low-dimensional index. We then make a clear representation of geographic-distributed features of spatial and temporal traffic flow patterns. In the spatial traffic pattern identification, the heat map of spatial pattern anomaly in the region shows almost identical results on weekday and weekend; if one detector suffers a higher anomaly degree, other detectors in the same intersection are more probable to be abnormal; an identical trend from AM peak, Noon, and PM peak can be found in different sub-regions. In the temporal traffic pattern identification, the occurrence of temporal abnormal places are quite random, and detectors in the same intersections may have quite different anomaly degrees from each other; temporal anomaly degrees in the same time period on different days usually have larger fluctuations and the higher value may indicate a bad traffic performance. In addition,
it is proved that there exists almost no correlation between spatial and temporal anomaly degrees. The combination of the two leads us to a better understanding of traffic flow pattern.

(2) In Chapter 4, we propose the Geo-Mobility method to study travel behavior and decode the travel motivations behind the abnormal directed travel.

We explore the characteristics of individual travel behavior especially the location geo-distribution, movement scale and the clustering features of undirected travel. We can conclude that: First, the travelers intend to stay within a certain range with a high probability. The travelers’ locations may assemble within several different clusters of different scales. Second, Locations show unique clustering features and spatial location differences within a cluster can be the results of the undirected travel. Third, in each location cluster for a single traveler, the visiting times of locations approximately follow a multivariate normal distribution across the geographic span. Based on these features, we propose a geo-mobility based method to capture the clustering features of traveler’s hourly locations, which can overcome the effects of the undirected travel. The abnormal travel behavior can then be detected. Finally, the abnormal behavior is further decoded by the tweet contents to find their travel motivations. 46.2% of these tweets posted in the abnormal locations contain keywords indicating major social activities. The travel motivation information extracted from historical data can be provided for the travelers for better route choice and travel schedule.

(3) In Chapter 5, we study the human mobility patterns of the travelers and decode the travel motivations behind the long-distance displacements.
First, we propose a new method to generalize the mobility features from high-resolution Twitter data and further unveil several important features of the trip displacements. Within 1-hour time window, our findings are quite different from that based on cellphone or GPS data: an ever-dominant portion of displacements lower than 100-m were found which accounts about for 74% of all displacement records and the portion of 0-m displacements even accounts about for 67%. This may be due to that the resolution of Twitter is much higher than those datasets in the previous studies. The Twitter displacements can be approximated by a Lévy Flight whose step size can follow heavy-tailed probability distribution. The $\beta$ value, which determines the fundamental mechanisms of the displacements, is much different from that of the previous studies in (Brockmann et al., 2006) and 1.65 in (Gonzalez et al., 2008). Second, we propose a method to study the travel motivations behind the displacements especially those long-distance ones. The Latent Dirichlet allocation (LDA) is employed to extract the topics from the tweet contents behind these long-distance displacements. Our study can be applied on two promising applications: individual travel motivation identification and travel pattern identification in a region towards a social event.

(4) In Chapter 6, the study of traffic accident identification is conducted based on both the traffic and social media data.

The methodology, we adopt, extract both the individual and paired token features for regression analysis and results with different features are presented and discussed. The prediction employs three important features: single token, paired token and traffic-related data to achieve a more accurate and effective on-site traffic accident detection. Our results show that paired tokens can capture the association rules inherent in the accident-related
tweets and efficiently increase the accuracy of the traffic accident detection. In comparison, the derived traffic-related information barely improves the results of accident prediction. The comparisons between the accident-related tweets and the traffic accident log show that: First, sometimes the tweet reports the traffic accident much faster than the traditional media, so detecting the online accident-related tweets can significantly reduce the response time. Second, tweets can capture those “mild” accidents that do not incur the attention of traffic police and this indicates the possibility of tweets detecting unreported accidents. Third, tweets can provide false alarms. Some tweets cannot give an accurate location of the accident site and more data sources should be further examined. The accident information extracted from social media may help the drivers to keep informed about the road conditions and detour the accident route in a real-time manner.

(5) In Chapter 7, the **Twitter concentration**, traffic surge and their inter correlation are studied.

First, a probability index is proposed to quantify the level of detector-based traffic surge in a large-scale road network. Second, an effective detection method is proposed to extract, filter and classify the traffic-related Twitter concentrations from a total collection of tweet posts. Third, a methodology is developed to evaluate the correlation between a specified tweet post and its surrounding traffic. Our experiments show that 77.4% of traffic-related Tweeter concentrations can be justified by local traffic surge. There are promising applications for this study: First, based on “Big Data”, it can precisely unveil the traffic patterns in a large road networks and even identify the anomaly traffic conditions. Second, this study can help traffic operators understand the cause of traffic surge and improve short-term prediction of traffic congestion (especially non-recurrent congestion) on roadways in
the future. Third, the Twitter concentration can broadcast the traffic-related events in a much more timely and quickly manner than traditional broadcasting media.

(6) In Chapter 8, an empirical study is conducted, as the first attempt, to infer the trip purpose by joining Connected Vehicles trajectories, land use data and social media data (Twitter).

We examine the tweet data, extract the trip trajectory data finding some empirical results as a preliminary step. According to the features of the tweet contents, we extract the tweet locations and label them with Google Place Types; our studies show that the tweets can report place names, which can be extracted by first searching for the proposition “at”, “in”, “on”, etc. and then manual examinations. Finally, we rank the tweet locations around a trip and infer the trip purposes. One can see that the tweets can reflect the social preferences and popularities of an activity location as people may tweet about the places at different frequencies and periods. Also, the tweet location category obtained from Google Place Types provide informative results for trip purpose inference study which prove the advantages of tweets by comparing with the land use data.

9.2 Future scope

Several important extensions are worth mentioning at the end of the dissertation and they will certainly bring about more possibilities in future research. The extensions may include the data analysis with higher dimensionality, event identification with more ground truth, travel behavior study with field investigations and many other possible applications:

(1) More historical data involved to increase the accuracy
In the compression-based traffic pattern identification in Chapter 3, applications in transportation data analysis still show certain limitations. First, the method works better on a large dataset. It is better to employ the method on the historical traffic data with long recording period. Future works can be more precise and meaningful with more historical data.

The human mobility study in Chapter 5 opens a window to access the human mobility features through Twitter displacement. It not only brings a mobility pattern result with higher resolution but also interpret the travel motivations behind the long-distance displacements. The results are based on 200 tweet users who have a relatively high tweeting frequency. The mobility pattern study especially that on a social event will be more accurate with more tweet users involved. Future studies will tackle the problems by incorporating more social media tools.

In the traffic accident identification in Chapter 6, one can see that an accurate automatic classification method for the social media data is worthy further investigations. With increasing coverage of social media, the automatic analysis of tweets can detect and even evaluate big social events and further their impacts on the travel behavior.

(2) More crowdsourcing data to unveil the travel information

In the social media study, we only consider the potentials of tweets, which already gives promising results. Further studies can include other social media data sources such as Facebook, WAZE, Square, etc.

In the travel behavior and human mobility study in Chapter 4 and 5, one limitation of this study is that we lack ground truth of people’s travel motivation or trip purpose. Future
studies shall consider this limitation by conducting supervised survey and experiments which can be taken as the ground truth.

In the Twitter concentration study, one can see that integrating social media data into the social event study opens up a wide range of possibilities for research in crowd-sourced transportation. The results show that social media data might be noisy and even unreliable, so solely relying on social media data is still not a perfect option. Further studies can focus on the data fusion of different data sources to better realize the purposes of other research such as traffic jam detection, traffic emergency evacuation (Asamoah, 2014), etc. The spatial-temporal features of traffic data are also worth studying for regional traffic operations. Note that our tweet data are geotagged. It would be an interesting extension to detect traffic events with non-geotagged tweets.

In the trip purpose identification study in Chapter 8, the features of the tweet locations should be further studied, and a supervised learning method can be employed for a more accurate prediction. This problem can be overcome by collecting the tweets and trip information from the same traveler.

(3) Practical applications for advanced traffic assistant

We can extend our studies to more practical areas by designing the traffic apps for driving assistance. The application can be the Android, or iOS apps that can be used in mobile platform. Similar applications like WAZE can report the traffic performance in a real time manner. In our dissertation, three future applications are worth discussing:

- Apps that can report the road accidents by automatically examining real-time tweet streams. This application can base on our traffic accident studies.
• Apps that can record the travel behavior of individuals and find those with similar mobility patterns for car sharing.

• Apps that can record the trip purposes and further provide for the travelers the important information of surrounding facilities.

(4) Social media interaction to impact the travel behavior

In this dissertation, we mainly studies the potentials of employing social media to extract useful traveler information. In fact, the social media is usually not a merely reflection of what people plan to do or what they have already done. Sometimes, the travelers may refer the social media information before they plan their trips. The social media can thus impact the travel behavior.

The interactions between travelers on social media can be a worthwhile future study. In this study, instead of merely extracting information from isolated tweet users, we need to find the interaction groups in which the individuals are influenced. Our initial examinations also prove that as the following two tweet posts.

• Mike just came back from Captain America 3 and told me its an awesome movie. I will go also.

• Nice dinners with Mom, Thank your for your advices, Cal.

These shows that the tweet users can influence each other. For a tweet user, by studying the tweets from his friends may possibly find information of what the tweet user plans to do next. This is a good topic given the increasing coverage of tweets.

(5) Integrating the results of our research

Integration the results and tackling the same problem from different perspectives can be fruitful in future research scope.
In the traffic pattern study in Chapter 3, the method provides insight into the interpretation of the multi-dimensional traffic-related data, and some research questions remain, especially in incident detection, network evaluation, and traffic management. The incident or event can possibly be validated by the accident and Twitter concentration studies in Chapter 6 and 7. As the geographic distribution of spatial traffic flow patterns remains almost the same in different time periods, there may exist room for improving planning or management of the road networks. Thus, this can be combined with the land use study in Chapter 8. This can be further employed to study the travel demands of a specific regions.

In the trip purpose inference study in Chapter 8, one can see that the tweets can seldom identify the home which is a shortcoming in labeling the trip purpose. This can be solved by studying the corresponding trip trajectories and land use information as in (Bohte and Maat, 2009).

In Twitter concentration study in Chapter 7, monitoring the social media data may deliver useful traffic event information, including traffic accident, traffic jam, road construction, etc. It will also be interesting to combine the spatial-temporal correlations between traffic patterns and Twitter concentrations in future research. Other research topics such as the ridesharing, driver classification (Cui et al., 2016) can also benefit the crowdsourced data.
REFERENCE

2015, Ann Arbor Geographic Information System, City of Ann Arbor.
Aarts, H., and A. Dijksterhuis, 2000, Habits as knowledge structures: automaticity in goal-directed behavior: Journal of personality and social psychology, v. 78, p. 53.
Aggarwal, C. C., and P. S. Yu, 2000, Finding generalized projected clusters in high dimensional spaces, v. 29, ACM.


Cui, Y., Q. He, and Q. Wang, 2016, Behavior-based Traveler Classification Using High-Resolution Connected Vehicles Trajectories and Land Use Data: Transportation Research Record, p. submitted.


DMR, 2016.


FHWA, 2000, Highway capacity manual, Washington, DC.


Griffin, T., and Y. Huang, 2005, A decision tree classification model to automate trip purpose derivation: The Proceedings of the ISCA 18th International Conference on Computer Applications in Industry and Engineering, p. 44-49.


Grünewald, P., 2005, A tutorial introduction to the minimum description length principle.


Lee, J. H., S. Gao, K. Janowicz, and K. G. Goulia, 2015, Can Twitter data be used to validate travel demand models?: IATBR 2015-WINDSOR.


Lin, L., Q. Wang, and A. Sadek, 2013, Short-term forecasting of traffic volume: evaluating models based on multiple data sets and data diagnosis measures: Transportation Research Record: Journal of the Transportation Research Board, p. 40-47.


Mochihashi, D., 2009, LDA, a latent dirichlet allocation package: ATR Spoken Language Communication Research Laboratories.


Ni, M., Q. He, and J. Gao, 2014, Using social media to predict traffic flow under special event conditions: The 93rd Annual Meeting of Transportation Research Board.


Obar, J. A., and S. S. Wildman, 2015, Social Media Definition and the Governance Challenge–An Introduction to the Special Issue: Available at SSRN 2663153.


Ozbay, K., and P. Kachroo, 1999, Incident management in intelligent transportation systems.


Ranks-NL, 2015, Default English stop words list.


Stambaugh, C., 2013, Social media and primary commercial service airports: Transportation Research Record: Journal of the Transportation Research Board, p. 76-86.


Statista, 2016.


Stopher, P., and C. Wilmot, 2000, Some new approaches to designing household travel surveys–time-use diaries and GPS: 79th Annual Meeting of the Transportation Research Board.


Treiber, M., and A. Kesting, 2012, Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns: Transportation research part C: emerging technologies, v. 21, p. 31-41.


Zhang, Z., Y. Cui, Q. He, M. Ni, and J. Gao, An Empirical Study of Trip Purpose Inference with Connected Vehicles Trajectories, Land Use Data, and Social Media Data: Transportation Research Record, p. submitted.

Zhang, Z., and Q. He, 2016, On-site Traffic Accident Detection with Both Social Media and Traffic Data: presented at 9th Triennial Symposium on Transportation Analysis (TRISTAN IX).

Zhang, Z., Q. He, J. Gou, and X. Li, 2016a, Extremity and Influential Factors Analysis on Travel Time of Emergency Vehicles: Transportation Research Record, p. submitted.


Zhang, Z., Q. He, and S. Zhu, 2016d, Exploring Travel Behavior with Social Media: An Empirical Study of Abnormal Movements Using High-Resolution Tweet Trajectory Data: submitted to this TRB annual meeting.

Zhang, Z., M. Ni, Q. He, J. Gao, J. Gou, and X. Li, 2016e, An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge: Transportation Research Record: Journal of the Transportation Research Board, v. 35, p. 36.
