Using Twitter for Next-Place Prediction, with an Application to Crime Prediction

---

A Thesis

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science (Systems and Information Engineering)

by

Mingjun Wang

Aug 2015

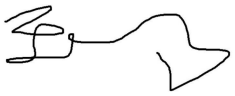
APPROVAL SHEET

The thesis

is submitted in partial fulfillment of the requirements

for the degree of

Master of Science



AUTHOR

The thesis has been read and approved by the examining committee:

Matthew S. Gerber

Advisor

Quanquan Gu

Hongning Wang

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

August

2015

Acknowledgment

I would like to thank my adviser Matt Gerber for his help and instruction for the past two years during my MS study. I have learnt a lot from him who is a good teacher and mentor to me. I would thank Prof. Wang and Prof. Gu, who have provided a lot of insightful suggestions to my research. I am deeply thankful to my parents Zhixiong and Xiaonan, for the love and support in the past years. Thank you every who's helped me succeed. Without your help I could not come this far.

Abstract

This research focuses on two problems. First, we investigate the prediction of social media users' spatial trajectories in the real world. Recent work on this task focused on utilizing cellular network traces and location-based social network services such as Foursquare, all of which emit structured geospatial information (e.g., cellular tower identifiers, GPS coordinates, and venue identifiers). Less attention has been paid to the rich textual content that users often publish in tandem with the structured information. From information in social media, we use two ways to define individual's movement pattern: the nearest venue type and the minimum distance to each venue type. We conclude that the combination of textual content together with a user's current location could be used to predict his or her movement pattern. We investigate methods of integrating textual content into existing next-place prediction models, and we demonstrate a significant improvement in next-place prediction compared to several baselines derived from published research. Second, we examine the correlation between routine activity extracted from next-place predictions and the occurrence of crimes in a major United States city, with the goal of aiding future research into automatic crime prediction. The result of next-place prediction with application in crime prediction will help policy making in police's law enforcement. The results support our hypothesis that people's movement patterns are correlated with crime rates. Then we have also built a classification solution to test how the movement trajectories help crime prediction. We extract the count of occupants who move to each venue type as features to demonstrate the movement trajectories. We build a classification solution to predict whether there is any crimes or not in a past time period. This preliminary classification model just utilize the density of people's movement to predict the happening of crimes, which need further investigation with more features like historical crime density.

Contents

Introduction

In our work, We first investigate how to automatically predicting users' spatial trajectories, a problem known as next-place prediction [1]. The main challenge in next-place prediction problem is to predict the next location of a mobile user given his or her current location. This work is built based on the assumption that the future events are determined by past events. Most existing work model next-place prediction problem as a classification problem using spatial and temporal features, like historical visiting frequencies. Intuitively, the sequence of locations that an individual visits will follow a well-established routine activity pattern, so that the next location could be predicted with transitions between different types of places and spatio-temporal characteristics of individuals' movement pattern. Our ultimate goal is to apply next-place prediction to model individuals' daily movement for predicting crimes with the help of social media. To understand individual's daily routine activity, the type of venues provide more information than raw geographical coordinates, which is one of the key factor to the happening of criminal activity[2]. Thus, we study the problem of next venue types people would visit instead of the latitude and longitude of an area. Traditional work in next-place prediction relies on mining individual's spatial historical trajectories[3, 4]. Recently, with the rapid growth of location based social network (LBSN), researchers have started to study capabilities of social media for next-place prediction [5, 6, 7, 8, 9, 10].

However, they only use historical visiting information to predict individuals' next location without considering textual content in social media, as there is not usually an overt connection between textual content and historical visiting trajectories. Also, people often disclose hints about their daily activity in social media [11], so it is conceivable that their movement pattern would be correlated with their social media posts contents. LBSN services such as Foursquare and Facebook Places allow users to "check in" at venues and broadcast this information to their social network, which allow us not only know the geographical coordinates of a user at a given time but also the exact type of places people go like a restaurant, airport and so on. Most recent work on next-place prediction has used the locations of these venues to predict users' next-place trajectories. We use geotagged tweets from Twitter to provide textual contents together with user's movement trajectories. It has been demonstrated that Twitter has strong potential in predicting and describing election results [12], natural disasters [13] and crime [14, 15]. Previous work in next-place prediction has largely ignored the textual content of social media posts, which we hypothesize can substantially improve the accuracy of next-place prediction models. Thus, our main hypothesis for next-place prediction problem using Twitter is as follows:

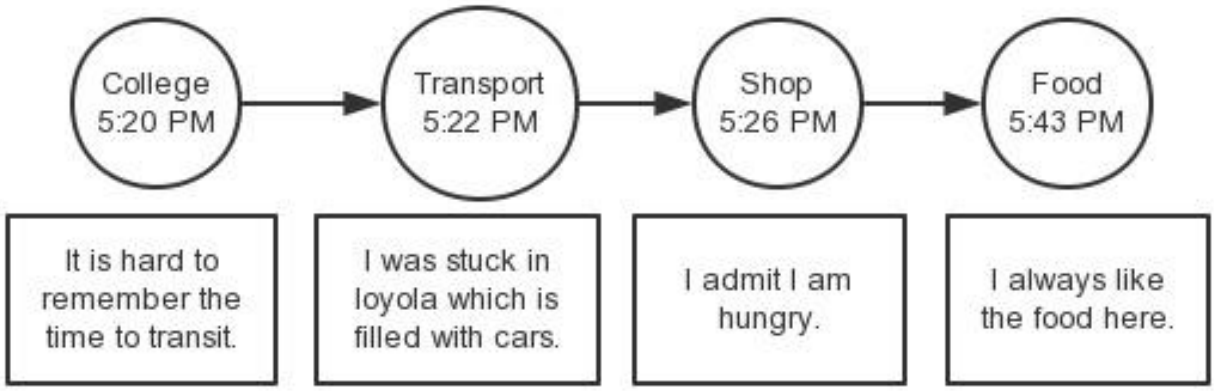


Figure 1.1: Venue Trajectory Example

- H1: An individual's future venue trajectory correlates with his or her historical tweets.

Incorporating historical textual content with individual's current information would help to model people's venue trajectories. Figure 1.1 shows an example of venue trajectory that demonstrate the correlation between venue trajectories and historical textual content,

The primary challenge in using Twitter for next-place prediction and crime prediction is that the textual content of tweets does not typically bear any overt connection with geospatial locations. Occasionally, a user will mention a specific address or business, or the user will attach a Foursquare check-in to their tweet. More often, indicators of future movement are left implicit. Consider the following tweets which demonstrate these cases:

- Mention of a specific address: Thanks to the "lady" at 3737 North Western Avenue, I'll never order Pete's Pizza again. And @GrubHub, done with you.

- Foursquare check-in: Check out ESPNChicago - #StateStreetStudio (190 N State St, at Lake St, Chicago) on @foursquare: http://t.co/PnFFbkTQ3m

- Implicit trajectory hint: @joshua_ocampoo: I'm hungry

The primary contribution of this research is to address posts such as the third one above (implicit). In this case, we are uncertain about the user's future spatial trajectory, but the textual content presents information that might imply movement of the user from his or her current location to a local dining establishment. To

be more specified, the geotagged tweets would provide two types of social information, the textual content and social relationships. We extract the social relationship with "@" symbols in textual content, so that a user's friend in social network will be grouped with his or her mentioning users. Specifically, our hypotheses for these two type social information are:

- The venue trajectory of a user's social network friends correlates with the user's own venue trajectory.
- A user's historical textual content correlates with his or her future venue trajectory.

In our work, we build text-enriched methods to improve the state-of-art in next-place prediction problem. Meanwhile, social ties are extracted with textual content by connecting different observations with mentions and replies extracted by @ symbols in tweets in this work. Intuitively, an individual's decision about which location they will visit will be influenced by other people's opinion. Therefore, we utilize social media information to build a classification solution for next-place prediction problem to model and describe people's routine activity.

Next-place prediction problem has a variety of potential applications in location-based advertisement, traffic planning and recommender services. For example, web search engines and location-based network services (LBSN) could provide sponsored advertisements related to the predicted next visiting locations together with search results or other service results. In our work, we aim at applying next-place prediction to crime prediction. Thus, we tend to build a connection between individuals' routine activities and localized criminal activities, each within urban environments. Crimes are more likely to happen at the space-time confluence of attackers, victims, and absence of protective elements as demonstrated in Routine Activity Theory[2, 16]. Thus, being able to predict individuals' movement patterns could be a useful aspect of effective crime prediction and policing. There are a lot of ways to describe and predict people's movement with the help of mobile devices and we will utilize the predicted result of our next-place prediction model. With the predicted result, we study the correlation between next-place predictions and the occurrence of crime. Specifically, our hypothesis is as follows:

- H2: Crime rates correlate with the density of users' movement trajectories in the same area.

Furthermore, with results from next-place prediction, we could understand and predict the relation between individual's routine activity with the occurrence of certain criminal activities. In our work for crime prediction, we first do this correlation analysis to study the relation between people's movement pattern and crime rate. Then, we also make a preliminary experiment on how to implement people's movement pattern to predict several most commonly happened crimes. Figure 1.2 shows an example of how we will get the correlation between the density of people's movement and crime rates. For example, the result of next-place

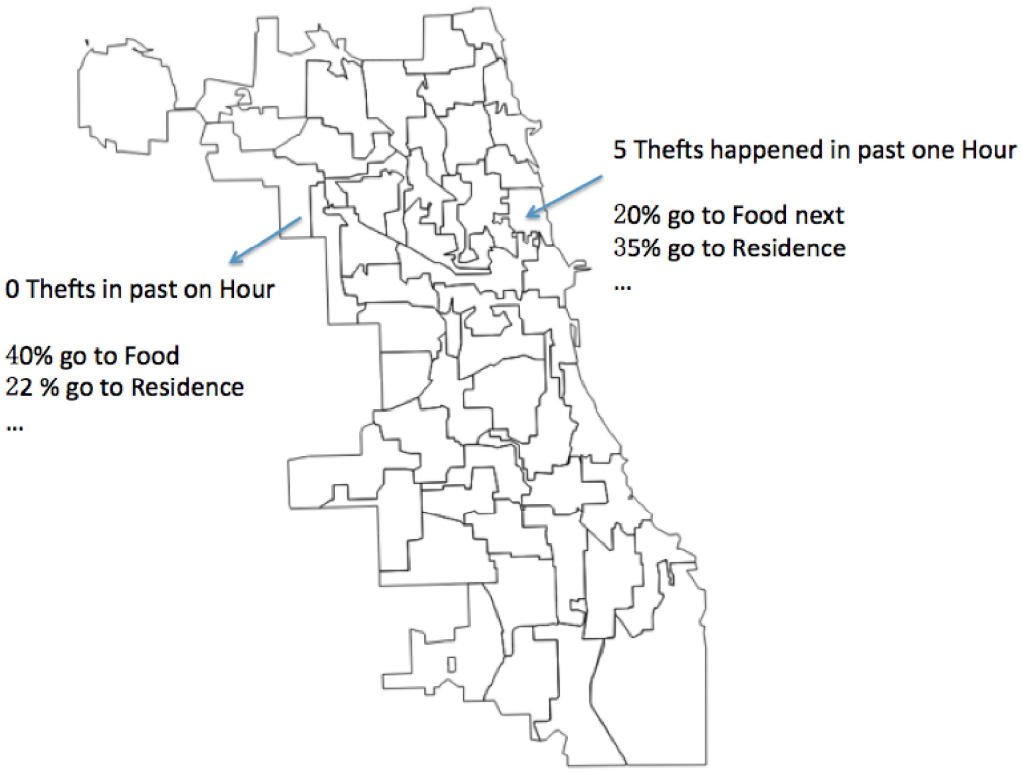


Figure 1.2: Correlation between Crime Rate and People's Movement Pattern Density

prediction problem shows that a user who post a tweet at Food venue will go to Residence venue as their next location. Then, we could group the geotagged tweets in a specified time window in the past in an area and get the crime rate in the same place and time period, like there will be 30% occupants in this area will go to Residence places and 7 incidents of Thefts. With this result, we would figure out whether there is any correlations between the count of occupants to different type of venues and crime rates.

The application of crime prediction using next-place prediction utilizing tweets context will both help the policy making in government's law enforcement and for commercial purpose. The prediction of crimes could help police to allocate appropriate guidance to the right place when potential victims and offenders are converge in space and time. For example, with the predicted result, we would probably know some potential victims and offenders will converge at late night during the movement from Food venue to Nightlife Spot venue. Then the police could send more guidance for the area with those movement which would help to prevent the happening of crimes. These applications could both help the government for policy making and protect local people. Meanwhile, social media information is easy to get within urban environment which make our methods are convenient to apply for next-place prediction problem. However, this method could not be directly applied to cities where GIS information simply does not exist or there is not a large population of

people using Twitter like China. Also, mining individual's movement pattern through social media messages will cause concerns of influence to personal privacy.

In summary, our contributions are mainly in two sides, which are improving next-place prediction using social media information and analyzing correlations between crimes and routine activities. For the next-place prediction problem, we present two models for text-enriched model. The first predicts the type of venue (e.g., restaurant or transport hub) the user will visit next. The second predicts how far the user will be from each type of venue (e.g., 100m from a restaurant and 3500m from a transport hub). We then test the correlation between predicted concentrations of users at these venue types and the occurrence of future crimes at such venues. Thus, our contribution is twofold:

- We develop and test a text-enriched model for next-place prediction based on social media posts.
- We formally test the correlation between next-place concentrations and the occurrence of actual future crimes in a large United States city.

Literature Review

In general, our work are mainly in two respects. We are aim at incorporating next-place prediction problem to crime predictions. Thus, we first go over previous works in next-place prediction problem. Then we will summary some recent work about crime prediction and introduce environmental crimes theories including Routine Activity Theory, which is the theoretical connection between the predicted result of next-place prediction and criminal activity.

2.1 Next-Place Prediction

There has been substantial research performed on various aspects of next-place prediction. For the past work in next-place prediction, we will first go over the response and applications of next-place prediction problem. Then, we will introduce different categories of next-place predictions problem based on the using of different types of predictors and algorithms.

2.1.1 Next-Place Responses and Applications

In general, there are two types of work in this area. First is the prediction of an individual's home location as only a small portion of users tend to directly post their home locations in social media [17]. The current work in home prediction uses content including locations, time stamps and social relations from social media

[17, 18, 19]. They explorer certain activity area like home or work place by mobile network data [5] through people's movement trajectory and frequencies. For example, they assume people would following similar routine activities in weekdays, so that people would visit their home and professional places regularly and spend most of them time in this two places, so that it is possible to extract his/she home locations. Second is the prediction of an individual's location at any time, which is the focus of our work. Traditional work aim at mining mobility features from a variety of mobile devices. Most work utilize mobile network data like location information from GPS or Wifi and they focus on movement trajectories [20]. This work will utilize simple, intuitive features with some mobile data-specific methodologies to evaluate the performance of people's real life movement pattern. More features like messages logs could be utilized through smart phones. With rapid growth of LBSN, more and more focus has been on information from social media. Some researchers [6] utilize a sequence of venues extracted from foursquare as individuals movement pattern. As our ultimate goal is to extract the occupants move to each type of places. Different from the previous work using each individual's movement pattern, we pay more attention to the count of occupants move to different type of places next.

With the result from next-place prediction, there are a variety of applications from next-place prediction like mobile advertising [21] and disaster relief [22]. For example, given a user's current location, we would predict his or her next locations, so that appropriate advertisement will be delivered to the right costumer who will visit this place. For now, there has been no work applying next-place prediction to model individuals' routine activity to predict crimes. Meanwhile, our work has two parts including the nearest venue type and the distance to each venue type. Most of previous work did not consider the distance to each type of place, while we built a regression model to predict the distance to each type of places to describe the physical environment.

2.1.2 Next-Place Predictors

There are a variety of predictors used in next-place prediction. Firstly, traditional methodology in next-place prediction is using cell phone data to mine user's spatial historical trajectories[3, 23, 24]. They build some classification solutions with spatial and temporal features extracted from mobile devices. In more recent works, researchers focused on rich contextual information from smart phone sensors [25]. There are a variety types of features like, temporal features, accelerometer features, call/short messages features and system features. The methodology is similar to previous work in analyzing the trajectories but the researchers are aim at getting more information regarding the behavior of each user. With the mobile data, the visiting frequency is regarded as a key factor in the problem. Secondly, with the rapid growth of LBSN, researchers

have used check-in patterns to predict the next check-in. From LBSN, researchers extract a sequence of venues which is the exact type of location instead of geography coordinates. In [4], researchers further investigated the problem by using a set of features describing the transition between different venue types. Moreover, some researcher measure and compare the similarity between different users in social media for next place prediction by collaborate filtering [7].

However, most of previous work only focus on the spatial temporal framework for venue trajectories without considering other context. They only consider the features like visiting frequencies, transition patterns, temporal effect and so on. Our work contribute this work with a new methodology utilizing rich textual context together with the venue trajectories. And we have also utilize the current location in a new way by the minimum distance to each venue type in the regression model which is not used in previous work.

2.1.3 Next-Place Algorithms

First of all, some prior works model next-place prediction as a classification problem using historical visiting information for training. They predict an individual's next-place based on sufficient amount of people's moving history[23]. Secondly, some work build Markov Model to consider the patterns in the data [24]. In order to consider the movement patterns, [26] extend the markov model to mobility markov chain and get a relatively good result. In most conditions, the prediction of check-in patterns could use the similar methods as methods using in mobile data. Thirdly, with the social media, some researchers apply some measures to extract the social ties in social media to predict the check-in pattern[6]. From the algorithms in social media, there are some work about the location recommend system, researchers propose to consider user's reviews and check-in history for location recommend[27]. Moreover, some researchers focus on the spatial temporal pattern [28, 29] to build the spatial temporal framework for building the model. In [30], researchers propose a logistic regression model and identified strongest predictor is the check-in frequency of the historical check-ins made by the user. For the nearest type, our classification solution has similar framework with a variety of features extracted from Twitter. We choose support vector machine as the classifier in our classification solution and fit a linear regression for the regression model.

2.2 Crime Prediction

Law enforcement require unravel the complex data to assist operational personnel in arresting offenders and directing crime prevention strategies. There are two major methodologies in crime prediction, one is Knowledge Discovery in Databases (KDD) technique to discover crimes and the other is environmental criminology which is aim at understanding the behavior of offenders. Some traditional work is using behavioral

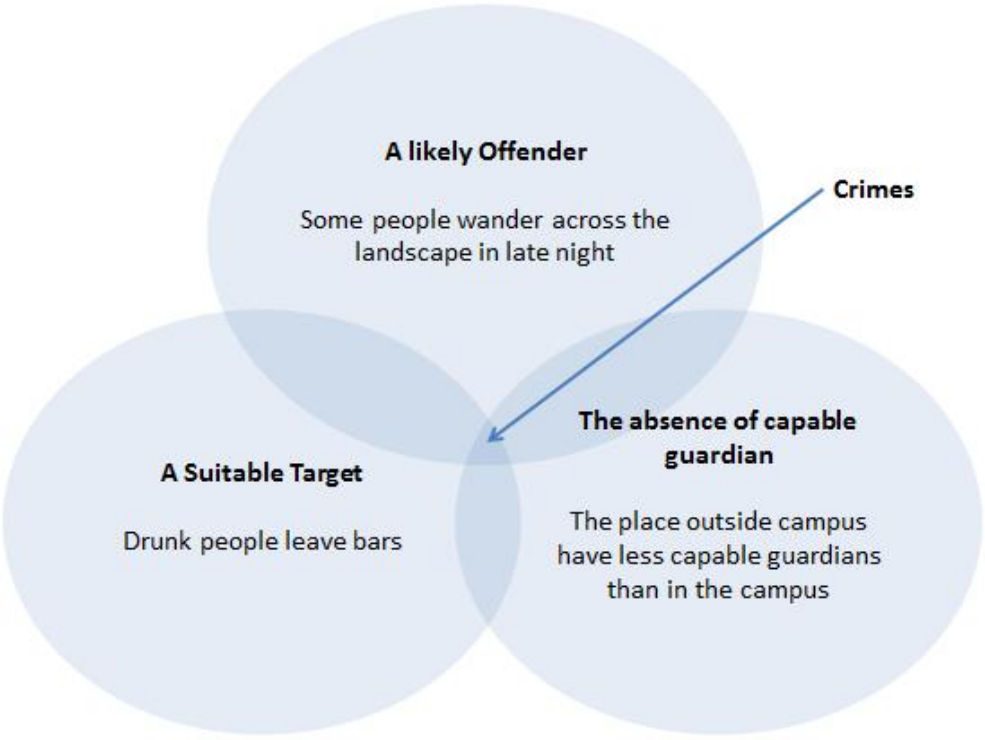


Figure 2.1: Graphical Model of Routine Activity Theory

profiling to find the patterns of unlawful activity, to predict the likely location and time of crimes, and to identify their perpetrators [31] which is an application as Knowledge Discovery in Databases (KDD) techniques.

Within environmental criminological theory, routine activities suggests that crime is likely to occur at the space-time confluence of offenders, targets, and the absence of capable guardians [16], as demonstrated in Figure 2.1. The Routine Activity Theory is the theoretical connection between next-place prediction and crimes in our work. For example, it demonstrated that individuals who leave their home are more exposed to potential offenders and are more likely to be the victims of criminal activities. Women who lived on campus were less likely to be victimized than those who did not. The study showed that women who lived on campus are surrounded by more capable guardians and are less like to be stalked. It assumes that crimes could be committed by anyone has the opportunity.

Also, there are some other methods in environmental criminological theory. The crime pattern theory tend to establish how people interact with their spatial environment and has three main notions: nodes, paths and edges [32]. The Rational Choice Perspective Theory focus upon the offender's decision making considering a few benefits and risk at a time [33]. The Awareness Theory has suggested that crime will be the

concentrated on the spatial element of crime including victim, offender, geo-temporal and legal to understand the behavior of offenders [34].

These days, the task of crime prediction and prevention often employ a variety of computer systems to identify and visualize areas of high crime which is known as "Hot-spots" [35]. Traditional hot-spot maps [36] produce retrospective visualizations of crimes, which can be predictive of future crime in cases where the occurrence of crime is stationary in space. Based on the plenty of historical crime information, researchers could identify hot spots where crimes are more likely have occurred in the past. These types of work has been implemented for years to help the police making to prevent crimes. These work illustrate that knowing occupants movement pattern in a time period in urban area will be likely to correlate with criminal activities. Thus, knowing people's movement could help to infer the appearance of suitable victims and potential offenders which is the focus of our work using next-place prediction for crime prediction. Crime prediction together with model people's routine activity will help law enforcement policy making.

Other research [14] integrates layers of geospatial information with historical crime records to improve these predictions. Also, some researchers employ twitter as the source for extracting events to predict crimes and use a major city in Chicago as a case study [15]. In our work, we provides preliminary evidence that these models could be further improved by including predicted concentrations of users at various venue types.

Data Preparation

We choose to use messages from social media for next-place prediction because users provide clues about their daily activity in social media. We use Twitter as the source for social messages to provide textual content and each user's current location. We prepare the data in three steps: data collection, tweet preprocessing, matching tweets with Foursquare venues. Thus, we extract geotagged tweets from Twitter and obtain typed venue locations from Foursquare. Considering tweets are informal and short, we filter the text contents by removing the stop words together with following parts-of-speech: determiner, postposition, coordinating conjunction, predeterminters, punctuation and numeral. Then we match geographical coordinates for each geotagged twees with typed venues to connect messages with the physical environment. For the next-place prediction problem, we propose two ways to describe the physical environment,

- Nearest venue type (categorical)
- Minimum distance to each venue type (continuous)

The nearest venue type is the location which is the closest to that individual which is more likely to be the type of place visited. So that the nearest venue type is discrete variable with ten levels for each venue

category. The minimum distance to each venue type is slightly different from the nearest venue type that is a continuous variable for each venue type unit with meters. For example, when a user post a tweet tagged with its geographical coordinates, we could know this user is probably in a food venue category which is the nearest venue type to this geotagged tweet and this geotagged tweet is 50m away to food venue and 1000m to event. Thus, we would know what type of a place this individual is visiting.

3.1 Tweets and Foursquare Venues Collection

We extracted geotagged tweets from Twitter and obtain typed venue locations from Foursquare. We collected geotagged tweets with textual contents, user ID and geographical coordinates of longitude and latitude through Twitter's streaming API. Following is an example of the context of the collected tweet,

- Time: 2014-01-01 00:01:09-05
- User ID: 439478296
- Geographical Coordinates: (-87.63846173, 41.88850092)
- Text: Addie and I get to celebrate New Year's Eve twice!!!!

All geotagged tweets are taken within the city boundary of Chicago, Illinois, USA in January 2014. We retained users who posted at least 20 tweets in this month. Thus, there are 1,233,076 tweets from 9,567 users in our data set. Figure 3.1 shows the every day count of tweets in January. The venues in Chicago are extracted from check-in histories on Foursquare, which include the following ten categories: Travel & Transport, Food, Residence, Outdoor & Recreation, Professional & Other Places, Arts & Entertainment, Nightlife Spot, College & University, Shop Services and Event. In total, there are 224,124 venues in Chicago. For example, we get a train station venue to describe a exact location. Then this train station venue will be grouped into the typed venue category Transportation. For each venue, we get what type of places it is together with its geographical coordinates. Following is an example of the context of the collected venue:

- Venue: Train Station
- Venue Type: Transportation
- Geographical Coordinates: (-87.6730120182037, 42.0189474398996)

For the computation convenience, we project the pair of longitude and latitude to meters in the urban area of Chicago. We use rgdal package in R to assign and transform this Coordinate Reference System. We

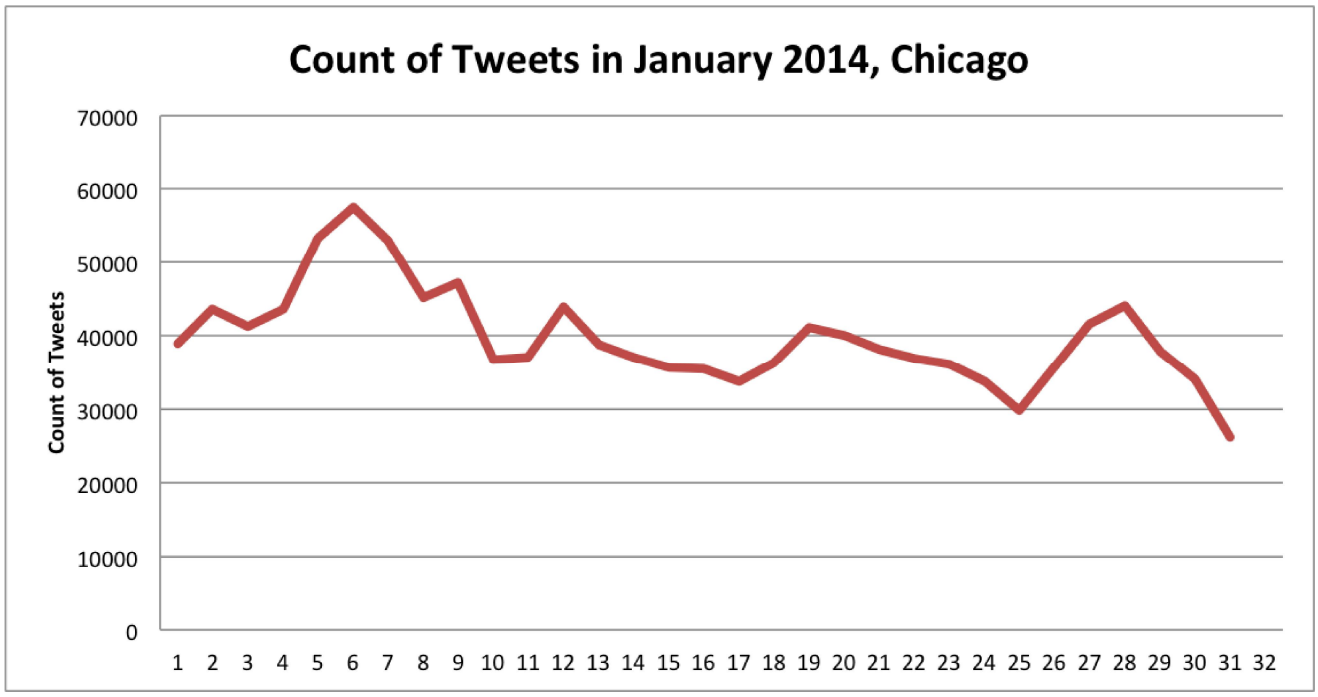


Figure 3.1: Count of Tweets in January 2014, Chicago

project the longitude and latitude pair points to Illinois East so that the calculated distance will be in meters. Thus, we will get the location in meters like the following example,

- Geographical Coordinates: (-87.63846173, 41.88850092)
- Location (meters): (357669.7, 579952)

In the following Table 3.1, we shows the count of each venue type in Chicago. Among all categories of venue types, the count of Residence venue is much larger than any other type of venues. And Event venue only takes a small portion in all type of venues. This result is consistent with our intuition that events like conference, festival and etc. only happened in a small range of the city so that it will have a small amount of venues and the Residence venues are more likely to be the place for people to post social media messages.

3.2 Tweet Preprocessing

For textual content, there are a lot meaningless parts in the textual contents in tweets which make tweets are difficult to use directly. The textual contents in tweets are informal and consist a lot of typos make it is challenging to extract appropriate features. We first apply part of speech tagging (POS) using Twitter POS model with the coarse 25-tag set from TweetNLP [37] which is a POS tool developed for the short and

Table 3.1: Count of Venue Types in Chicago

Venue Types	Count
Transport	21815
Food	20501
Residence	74310
Recreation	9931
Professional	39258
Entertain	8310
Nightlife	11312
University	4459
Shop	33873
Event	355

informal contents like tweets. We utilize Tweet NLP to tokenize the textual content and tag the tokens. For example, getting the following tweet text, we will get a list of tags, probabilities and tokens,

- Raw tweet: @Bmfayy dude he didn't play up to his potential I mean he got the job done but if they play like that against the 49ers they will lose
- Tags: @ N O V V T P D N O V O V D N V & P O V P O P D ∧ O V V
- Confidence: 0.9991 0.8033 0.9908 0.9999 0.9999 0.9030 0.9991 0.9969 0.9003 0.9922 0.9960 0.9984 0.9998 0.9996 0.9846 0.9586 0.9847 0.9961 0.9961 0.9988 0.8832 0.9294 0.9977 0.9993 0.7797 0.9657 0.9996 0.9977
- Tokens: @Bmfayy dude he didn't play up to his potential I mean he got the job done but if they play like that against the 49ers they will lose

Then, with the above result, we removed determiners, postpositions, coordinating conjunctions, predeterminers, punctuation, and numerals. Considering textual contents in tweets are informal, these removed tokens also included a lot of misspelling words which are probably be some typos only appear a small amount of times. Then we also removed stop words from a built-in stop words list in one Natural Language Toolkit [38]. Then we only keep a list of tokens for each geotagged tweets.

Also, we extract social relationship from textual contents. The "@" symbol with user name in the body of tweets stand for mentions and replies. As showed in the above example, the tweet will mention a user through making the user name following "@" symbol. A user will get a notification when he/she was mentioned by other users. Using the POS tool, we collect these mentions information for each tweet through extracting @username. For each user, they have a user ID and a screen name.

- User ID: 617396479
- Screen Name: lovee_esmeralda

Table 3.2: Distance between each geotagged tweet and their nearest venue (m)

Average	Standard Deviation	Max	Min
58.43	60.18	2768.47	0

For each screen name following "@" symbol, we match it with its user ID and then connect it to this user's current tweet. We assume there will be some relationship between users who will mention each other.

3.3 Matching Tweets with Foursquare Venues

To fill the gap between social messages and the physical environment, we matched each observed geotagged tweet with a type of venue to describe the tweets' surroundings and calculate the distance between the tweet and each venue type. For the geotagged tweets and location of venues, we first project the longitude/latitude pairs to meters within the city boundary of Chicago as described in the above section, so that the distance is in meters. Through calculating the distance between tweets and each specified venue, we first identified the nearest venue within each type (e.g., nearest Event, nearest Nightlife Spot, etc.) and record the distance to each venue type. Then we identified the closest venue among those venue types. These values constitute the responses for our two next-place prediction problems (distance to all venue types, and nearest venue, respectively). Table 3.2 shows that the distance between each geotagged tweet and its nearest venue. As showed in the table, the nearest venue is very close to some geotagged tweet which means this venue is more likely to be the place this individual is visiting. Meanwhile, some venues are far away from the posted tweets, so that the nearest venue probably not be the perfect description of the physical environment. Thus, we could also utilize the nearest distance within all categories of venue type. For example, considering the following geotagged tweet as one example,

- User ID: 1531861880
- Location: (358597.499911395, 579767.2494029)
- Raw Tweet: Happy New Year Tempa people!!!

For this specified geotagged tweet, we will calculate its distance in meters to all typed venue as following,

- Minimum Distance to Each Venue Type in Meters
- Arts & Entertainment : 23
- College & University : 43
- Event : 57

- Food : 15
- Nightlife Spot : 42
- Outdoors & Recreation : 50
- Professional & Other Places : 6
- Residence : 83
- Shop & Service : 40
- Travel & Transport : 34

With the distance to each venue type, we will also get the nearest venue type which is Professional & Other Places as the above example. As this matching step is very time consuming, we are using High Performance Computing platform Rivanna in this step.

Next-Place Prediction Problem

In this section, we present our approaches for incorporating textual content into next-place prediction models. When an individual posts a tweet, we define the next-place prediction problem as the task of predicting the location where this individual with post his or her next tweet. The locations in our next-place prediction problem are defined as a sequence of venues assigned to each geotagged tweet. We formulate the next-place prediction problem in two ways: predicting the nearest venue type and predicting distances to each type of venue.

4.1 Text-Enriched Classification Model

To incorporate textual content for next-place prediction, our text-enriched classification model predicts each user's next venue. The next venue is the venue nearest to the location where the user posted his or her next tweet. There are two types of movement for each user given his or her current location. One is stay in the same location so that the user will post next tweet with the same nearest venue. The other is move to a new location so that the user's next nearest venue will be a new venue different with the current location. Thus, we build our model in two parts. First is a binary classification model to determine whether an individual will maintain the current nearest venue type or move such that a different venue type become the nearest. Second, for users who will transit to a new venue type, we build a multivariate classification model to predict

the venue type that will become nearest. The general forms of these models are:

$$Step1 : P(c_{n+1} = c_n|\chi) = F(f_1, f_2, ..., f_n),$$

$$Step2 : P(c_{n+1} = v|\chi) = F(f_1, f_2, ..., f_n)$$

where χ is the user's historical visiting history $\chi = \{c_1, c_2, ...c_n\}$, c_{n+1} is next venue. For each c_i in χ, c_i is a venue type, rather than a particular venue instance. Predictor variables $f_1, f_2, ..., f_n$ are the features extracted from a user's historical tweets including historical venue trajectories and tweets associated with the user. We model each of the above classification problems with linear support vector machines implemented by LibLinear [39]. LibLinear is an open source library for large scale sparse linear classification which is much more efficient than most previous applications of support vector machine on large sparse data sets. We directly use LibLinear with its command-line tools for the learning task. We use the default classifier L2-SVM for the following training tasks.

With different sets of features, we have studied two classification models, which are the Text-Enriched Model and the Text-Enriched with @-link Model.

4.1.1 Text-Enriched Model

We extract features from the textual content of a user's tweets and the locations of these tweets to build the Text-Enriched Model.

- Hypothesis: Individuals' historical textual content and his or her current location correlates with his or her future venue trajectory.

Figure 4.1 shows a specified example of Text-Enriched Model, given a user's current location and social media messages, we could figure out his or her next visiting place. Under this hypothesis, f_1 the current venue type, $f_2, f_3, ..., f_n$ are TF-IDF features from textual content of the user's recent tweets so that the general forms of the models become:,

$$Step1 : P(c_{n+1} = c_n|\chi) = F(c_n, tfidf(t_n)),$$

$$Step2 : P(c_{n+1} = v|\chi) = F(c_n, tfidf(t_n)),$$

where χ is defined as the historical visiting history, c_{n+1} is next venue, t_n is the tweet posted in current location, and $tfidf(t_n)$ is the set of features extracted from the current tweet's textual content. Features extracted from textual contents are represented with Term Frequency Inverted Document Frequency (TF-IDF) as a vector space model, where the IDF component is calculated from all historical tweets which will be the

Figure 4.1: Text-Enriched Model Example

tweets in training set in our data. The TF-IDF is calculated as following forms,

$$tf(word, tweet) = 1 + log\ f(word, tweet), or\ 0\ if\ f(word, tweet)\ is\ zero,$$

$$idf(word, historical\ tweets) = log \frac{Total\ Number\ of\ Tweets\ in\ Corpus}{1 + Numer\ of\ Documents\ where\ this\ word\ appears}$$

$$tfidf(word, tweet, historical\ tweets) = tf(word, tweet) \cdot idf(word, historical\ tweets)$$

For each token remaining after tweet preprocessing, we calculate a TF-IDF score for the token in each tweet. For the current venue, we translate the categorical predictor variable for current venues into 10 variables each with two levels. Thus, we are using a set of numerical variables for textual contents and dummy coding with ten levels for current location.

4.1.2 Text-Enriched with @-link Model

The Text-Enriched with @-link Model extends the Text-Enriched Model with features extracted from tweets that mention the current user. This model starts with the feature set described above for the Text-Enriched Model, and it adds to this set features derived from the following hypothesis:

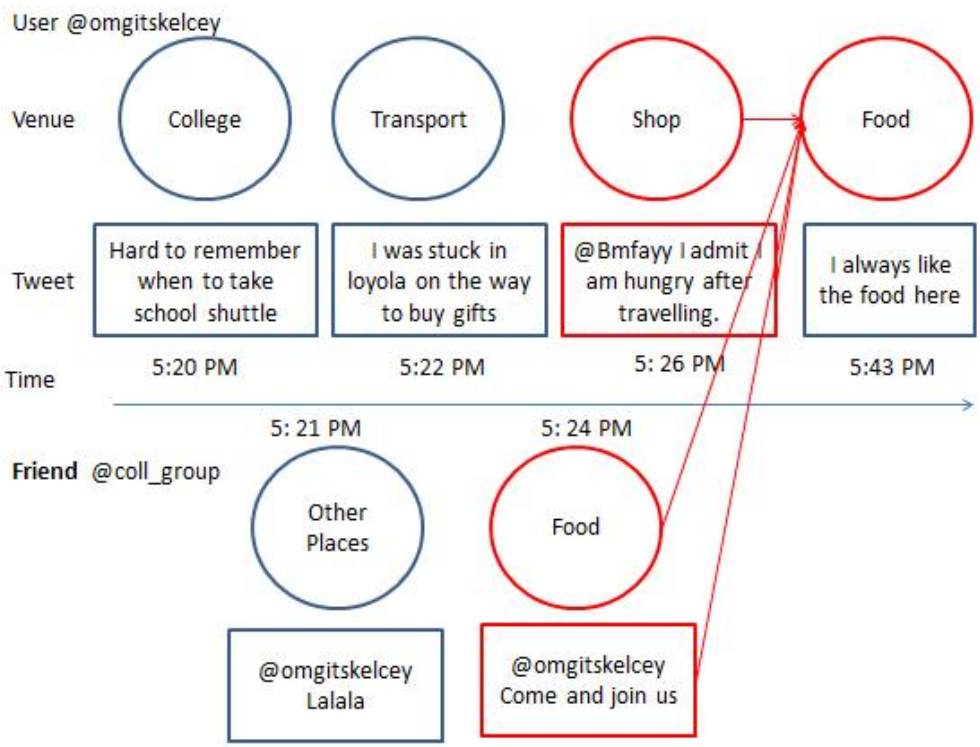


Figure 4.2: Text-Enriched with @-link Model Example

- Hypothesis: The current user's next location will correlate with the location of other users who have recently mentioned the current user.

Intuitively, users are prone to traveling to locations where their friends are located. We use the @-link in tweets to stand for such relation. Beyond a user's own postings, the user is probably mentioned by some other users as well which contains a lot of useful information. Figure 4.2 shows a specified example of Text-Enriched with @-link Model. As showed in the example, there will be a screen name following "@" symbol. We will match this screen name to its corresponding user ID.

For each tweet, there are two groups of features. The first is the same that used by the Text-Enriched Model, which includes current location and the TF-IDF features. This set of features are about the user's own information. Features from the second group are extracted from other tweets that mention the current user, so that features will capture the current user's location as well as the location of users in his or her social network. Thus, these features include two sets of TF-IDF features with the same size of tokens and two sets of dummy variables for the categorical variables stand for the user and the user's friend's location. If this user has not been mentioned by any other user, this geotagged tweet will only have exact the same feature as described in the Text-Enriched Model and all features in the second set of features will be assigned to be 0.

More formally, f_1 is the current user's nearest venue type, $f_2, f_3, ..., f_n$ are TF-IDF features from this user's textual content, $f_{n+1}, ...f_{2n}$ are venue type and TF-IDF features extracted from recent tweets that mention the current user. The general form of this model as follows,

$$Step1 : P(c_{n+1} = c_n|\chi) = F(c_n, tfidf(t_n), m(u)),$$

$$Step2 : P(c_{n+1} = v|\chi) = F(c_n, tfidf(t_n), m(u)),$$

where $m(u)$ is set of features extracted from most recent tweets that mention the current user,

$$m(u) = (c_n(u), tfidf(t_n(u))$$

Indeed, the user will probably be mentioned by many other users and we only take the most recent mentioning in our model.

4.2 Text-Retrieval Model

Intuitively, the textual content of tweets posted in proximity of the same venue type should share similar textual content. Thus, it should be possible to retrieve tweets corresponding to a venue type based on the textual content. For example, in some bars, people are more likely to post tweets related to the discussion of some football games. Thus, in the future, such tweets will be regarded as related to some places including this type of bars. Also, if the user mentions food in his or her tweet, this would presumably match tweets from food establishments, leading to a prediction of Food as this user's next place.

We build a collection of documents, one for each venue instance in the city. Assuming that the tweet will be posted in its nearest venue, each document is the set of all the historical tweets which are the tweets in training set in our experiment. We consider a user's current tweet as a query against this document collection. Thus, given a utility function for the query and documents, we could get scores for all the documents for each tweet. Then we could rank all the documents based on the similarity among their textual content. In the ranking list for each tweet, the top-ranked document with respect to this query might be the user's targeted next place. We utilize a ranking function to rank matching typed venue's documents according to their relevance to a give query tweet. As we only consider the textual content without any other context for the social media message, it is hard to say whether the textual content is correlated with its current location or next location. We will examine and compare the result for both current location and next location.

We implemented the above ideas with a simple vector space model of individual tweets (queries) and documents derived from tweets posted at particular venues. We used the BM25 [40] query model, which is defined as follows:

$$score(q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_i + 1)}{f(q_i, D) + k_i \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

q is the textual content of a tweet, D is the collections of historical tweets in one venue type. We will get a score for the tweet and the document for each specified venue. For any words q_i in q, we get $f(q_i, D)$ as term frequency in the document and $IDF(q_i)$ is inverse document frequency for the query term. $\mid D \mid$ is the length of document D, $avggl$ is the average document length in all venues. k_1 and b are free parameters. Without further optimization, we set $k_1 = 1.5$ and $b = 0.75$ [41]. The type of venue with the highest score will be the output of the Text-Retrieval Model.

4.3 Text-Enriched Regression Model

The features and formulation of the Text-Enriched Regression Model are similar to those of the Text-Enriched Classification Model, except that the response is vector of continuous distances, one for each venue type. For some tweets, the nearest venue is still not close to user's current location. Thus, using the vector of distances to each venue type could be a better way to describe the physical environment. The general form of the regression model is,

$$y = F(c_n, tfidf(t_n), m(u))\beta + \epsilon,$$

where,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{10} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{10} \end{pmatrix},$$

$$F(c_n, tfidf(t_n), m(u)) = \begin{pmatrix} F_1^T \\ F_2^T \\ \vdots \\ F_{10}^T \end{pmatrix}$$

The F_i^T are feature vectors capturing user's nearest distance to each venue type and features from textual content. c_n is a vector including the current minimum distance to each venue type. The other textual features will be the same as the Text-Enriched Classification model. Here we utilize linear regression for this regression

problem instead of support vector machine. We build two regression models. First is a main effect model that uses all of these features, including the current minimum distance to each venue type and the textual features. Second is the model with interaction terms among the distance to each venue type with pairwise products representing pairwise-interactions among these distances.

4.4 Baseline Models

To evaluate the performance of our work for next-venue prediction, we built baseline models for our work with nearest venue type and distance to each typed venue separately.

4.4.1 Nearest Venue Type Prediction

For the classification-based approaches for next-venue prediction, we built baseline models chosen from state-of-the-art published research:

Most Frequent Check-in Model[30]

[30] has showed that check-in frequency is the strongest predictor in predicting users' next locations which formulates our first baseline model. Considering each user's check-in history, the model generates a probabilistic distribution on venue types based on the visiting history. For the next location v, the probability is defined as,

$$P(c_{n+1} = v|\chi) = \frac{\#\ check-ins\ to\ v}{\#\ total\ check-ins}$$

The next nearest venue is the venue with the highest probability considering this user's historical visiting patterns. This method is consistent with our intuition that each user is more likely to visit in some certain routine activities pattern so that he/she probably visit similar places in the future.

Markov Model

Gambs et al. propose to address the problem by developing a Markov model to incorporate the k previously visited locations [26]. They address next-place prediction problem based on the observations of his mobility behavior over some period of time and the recent locations that he has visited. They utilize Mobility Markov Chain (MMC) in order to incorporate the n previous visited locations. The probability is defined as,

$$P(c_{n+1} = v|\chi) = \frac{Count\ of\ Same\ Visiting\ Pattern\ in\ Previous\ k\ Visited\ Locations}{Total\ Count\ of\ Visiting\ Venues}$$

Given all the historical visiting venues, the Markov model will calculate how many visiting records will have the same pattern. The output of this model is the typed venue with highest probability given the visiting location in a recent period. The Markov Model with order 2 performed the best in their experiments. In our experiments, we take both order-1 and order-2 as baseline models and the Most Frequent Check-in Model is the Order-0 Markov Model. Different from the Most Frequent Check-in Model, the Markov Model not only consider that the user will visit similar place to the past but also will be influenced by some more recent movement patterns.

Classification Model with Historical Visiting Information

In [42], multiple algorithms were compared in the prediction of next-place with mobility data. They build a classification solution with a variety choices of features. The authors showed that the visiting frequency to each venue extracted from a user's historical visiting information is the most significant predictor in the next-place prediction problem. Thus, this classification solution is defined as,

$$P(c_{n+1} = v|\chi) = F(f(v_1), f(v_2), ...f(v_{10}))$$

Given the historical visiting information χ, we could extract the count of visiting to each typed venue with $f(v_1), ...f(v_{10})$. Thus, we built a baseline classification model with features including visiting frequency to each venue type.

4.4.2 Minimum Distance to Typed Venues Prediction

Most traditional work in next-place prediction is focus on what location people will visit, so they did not concern much about the real distance to this location or how far away an individual is from that venue. Since the regression approach to distance prediction is new, a previously developed baseline does not exist. As a baseline for the regression models, we used the average distance from each venue type as the baseline prediction. We computed the average distance in the following ways: We first calculate the distance to each venue type for all the historical tweets. Then we get the average distance from the average distances in all historical tweets. Thus, for each typed venue, we have a distance as the baseline for the distance to the user's

location.

Next-Place Prediction Results and Discussion

We measure the performance of all models described above as follows. For the prediction of nearest venue type, we define the prediction accuracy to be the ratio of the number of correct predictions and the total number of predictions,

$$PredictionAccuracy = \frac{Number\ of\ Correct\ Predcitions}{Total\ Number\ of\ Predictions}$$

We used Mean Squared Error (MSE) as the performance metric for the venue-distance regression models, which is defined as,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_n - Y_n)^2$$

$\hat{Y}_n$ is the n predicted result and Y is the observed result. As the distance is calculated in meters, the measurement unit for this regression solution is meter.

5.1 Experiments and Results

To build the train set and test set for all of our experiments. We first convert the time stamp with each geotagged tweet to count of seconds starting from 2014-01-01 00:00:00-00. For example, 2014-01-02 16:34:25-00 will be converted to 146065. Thus, we use first 20 days in January which are the geotagged tweets posted in the first 1728000 seconds in 2014. The test test will be the final 11 days which are the time larger than 1728000 seconds in January. The result for the prediction of nearest venue type is showed in table II. We used bootstrapping to obtain the 95% confidence interval for each accuracy score. We train the model in the first 21 days first. Then we randomly generating the data with the original test set data by random sampling with replacement and get the prediction accuracy for the middle 95% to be the confidence interval of the result. In our experiments, the Text-Enriched Model and the Text-Enriched with @-link Model produce similar results. They both have better performance than the state-of-the-art baselines that we implemented as showed in Table 5.1.

Table 5.1: Results for next-venue-type classification. Models with * are Baseline Models

Models	Prediction Accuracy (95% confidence interval)
Most Frequent Check-in Model*	(0.5886,0.5915)
Order-1 Markov Model*	(0.5575, 0.5758)
Order-2 Markov Model*	(0.5282, 0.5661)
Classification model with historical visiting information*	(0.6333, 0.6476)
Text-enriched Model	(0.7122, 0.7158)
Text-enriched with @-link model	(0.7098, 0.7130)

Table 5.2: Text-Retrieval Model Prediction Accuracy

Current Visiting Venue	Next Visiting Venue
0.1241	0.1321

In Table 5.2, we show the result of Text-Retrieval Model. With the user's generated content, we retrieve a venue through the Text-Retrieval Model. We evaluated the accuracy of this model to determine whether the textual content of tweets correlates more with a user's current location or next location, so we get both the prediction accuracy to current visiting venue and next visiting venue.

We used Mean Squared Error (MSE) as the performance metric for the venue-distance regression models. The venue type with better result in baseline models will be more evenly located in the city so that the average distance will be more close to the distance between each geotagged tweet and this venue, as showed in Table 5.3.

Table 5.3: Regression Model (MSE)

Venue Types	MSE (Baseline)	MSE (Main Effects)	MSE (interaction)
Transport	37365	10460	10262
Food	30433	12073	11795
Residence	26861	9831	9650
Recreation	30091	10665	10416
Professional	16127	8693	8450
Entertain	85713	18727	18443
Nightlife	40084	12482	12148
University	131126	26570	26404
Shop	20734	9233	9077
Event	745612	141189	140985

5.2 Discussions

For the nearest venue type, the results in Table 5.1 support our hypothesis (H1) that the content of geotagged tweets correlates with users' future venue trajectories. Although most users do not overtly mention their intent to move to a new venue type, they do reveal clues about this movement in the words of their messages. Thus, incorporating textual content to the model will improve the performance of next-venue prediction. It is possible to imply an individuals' location through knowing an individual's current location together with his or her social media messages. This result suggest that police could probably know the movement patterns in urban area through monitoring the social media messages in the city. This could be applied in for commercial purpose like location based advertisement. By investigating a user's social media content and knowing the next-place this user will go in the future, the specified advertisement will be deployed in the right time at right place. For example, we could probably know a user who like to eat sushi and will go to Food venue as their next location, so that the advertisement about Japanese food would probably be delivered to that user. Also, among all the baseline models, we should also notice that knowing the visiting frequencies to all categories of venues. Then, considering the social relationship, extracting information pertaining to social relationships from @ mentions in tweets did not improve the text-enriched model, suggesting that people's movement plans may not be influenced by the textual content of messages posted by their network peers.

Table 5.2 shows the result for Text-Retrieval Model. We find that the textual content of users' tweets has more prediction accuracy for next venue type, thus they probably correlate more strongly with their next venue type instead of their current venue type. In other words, people are more likely to share some information about their future activity rather than their current location. These results suggest that the Text-Enriched Model of next-venue classification might be improved in the future by incorporating a measure of similarity between a user's current tweet with tweets that have been posted from various venues. Comparing the result of Text-Retrieval Model with the classification solution, we found that combining the textual content with its current location will be much more powerful in predicting individual's next location.

The result for the distance to each venue type is showed in Table 5.3. The predicted result are evaluated through MSE, which is the second moment of the error and considers both the variance and bias. To most venue types, our model shows much better result compared to the baseline model that he predicted error will be hundreds of meters. We also found that the University & College and Event venue get worst in the result. The Event venue (e.g., conference or festival) and Universities are normally concentrated in a small areas of the city, where do not have these venues near posted geotagged tweets some time. Thus, the distance to this two venue types will be very skewed which cause the performance worse than other venue types. Both the main effects regression model and regression model with interaction terms show improved performance in

predicting the distances to venue types compared to the baseline model (results shown in Table IV). These results showed that the distances to each venue will varied with different geotagged tweets, so the average distance could not be good enough to illustrate the environment. Both regression model achieves superior distance predictions for most venue types, with Event being an exception. The location and textual content of current location could be used to predict the future locations as we could get better result than baselines. People's distance to the venues around it. The model with interaction terms are slightly better than the main effects model in all venue types. Also, there probably be some correlations within different venues which also help our model get improved performances.

In summary, the result support our hypothesis for next-place prediction problem which suggests that our model with features extracted from social media messages could get significant improved performance as showed in Table 5.1 and Table 5.3. We also notice that the correlation between textual content itself and next visiting locations as showed in Table 5.2. These results suggest that user generated content is informative to user's movement plan and correlate to his or her venue trajectories. Such result shows that we could use social media content to describe and predict people's movement plan. Extracting people's movement pattern have a lot of potential commercial application. The police could utilize this result for crime prediction which will help policy making in law enforcement. For example, we could probably know more people will move to nightlife spot venue. Thus, the police will come to prevent crimes when there are increasing routine activities. This application for police making will be discussed in more details in the following chapter of analyzing the correlation between crimes and next-place prediction result. Meanwhile, predict user's venue movement through their social context will also raise some privacy problems in the future.

Correlation Analysis of Crimes and Next-place Predictions

We hypothesized (H2) that crime counts would correlate with the predicted concentration of users at various venue types. In the following sections, we first do a correlation analysis to examine the relation between crime rates and next venue type occupancy. The methodology section presents analysis and results for this hypothesis. Following that, we will show the result and demonstrate some interesting findings in the Result and Discussion section. Then we build a preliminary classification solution using features extracted from people's movement pattern based on the predicted result of next-place prediction problem.

6.1 Correlation Analysis Methodology

Users' movement patterns are defined as the predicted concentration of users at each venue type. When an individual posts one tweet, his or her movement pattern is the movement from the current location to the next location where a tweet is posted. For example, when a user post a tweet, we would predict his or her next visiting venue. Then we will group the users who are predicted to visit each venue in a certain time period and within some areas. Thus, we could get some results like there are 10 occupants in the past one hour will visit Night Life Spot venue as their next location. We assume such result will demonstrate people's routine movement activity pattern within the urban areas.

We utilize the output of the next-place classifier to determine the concentration of users at their next venue types. For each venue type, we define the predicted occupants as follows:

$$Pre(p) = \{c_1(p), c_2(p), ...c_{10}(p)\}$$

p is a spatial point in a grid of evenly spaced 2000-meter squares across the city, $c_i(p)$ is the count of predicted occupants for one of ten venue types within a 2,000 meter square across the city in the past one hour. The predicted occupants are defined as count of users who will visit each venue type. We calculate all the points in the city every hour in January, 2015 in Chicago.

For crimes, we collected all records from Jan 1, 2014 - Jan 31, 2014 from the Chicago Data Portal. There are 25 crime types including 19,691 instances in total. Table 6.1 shows the count of crimes in each crime type in the city from January 2014. Some crimes like gambling, stalking and so on only have several observations in the crimes records. For the purposes of our study, we retained crime types with at least 1,000 instances. Similar to the next-place concentrations described above, we count the frequency of each crime type within a grid of 2000-meter squares that cover the city, on an hourly basis. These squares are identical to the squares used for next-place concentrations, allowing us to calculate basic correlation statistics, as follows.

With the paired counts of crimes and users' movements to each venue type, we calculate the correlation between crime count and venue type occupancy. For example, we have two variable which are count of Thefts every every for each grid and the count of occupants to Food venue for the same time period and area. Then we will calculate the linear correlation (dependence) between these two variables to evaluate the correlation between crime rate and occupants venue type. Actually, we will calculate this correlation score for each crime type with each venue type. For the correlation score function, we used Pearson's Product Moment Correlation:

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Table 6.1: Count of Crimes in Chicago, January 2014

Crime Type	Count
Gambling	1
Stalking	7
Intimidation	10
Kidnapping	20
Homicide	20
Arson	21
Liquor Law Violation	31
Sex Offense	65
Prostitute	68
Crime Sexual Assault	81
Interference with the Public Officer	90
Public Peace Violation	169
Weapon Violation	201
Offense Involving Children	230
Criminal Trespass	575
Robbery	797
Motor Cycle Violation	806
Burglary	1134
Assault	1036
Deceptive Practice	1138
Other Offense	1407
Criminal Damage	1789
Narcotics	2222
Battery	3335
Theft	4438

We calculated this correlation for ground-truth next-place trajectories as well as the predicted next-place trajectories from our Text-Enriched Classification Model. Also, we used p-value via the asymptotic t approximation to evaluate the significant of the correlation relation. The null hypothesis is that there is no correlation between crime count and venue type occupancy.

6.2 Results and Discussion

The results of our analysis are shown in Tables 6.2 through 6.9, where each table shows the correlation between crime counts and predicted concentrations of users at the various venue types. Across the tables, we see that many crime types have significant correlations with predicted venue concentrations. For example, the occurrence of burglaries is positively correlated with the transition of users to shopping & service destinations. At this point in our work, we are not clear on the causal mechanism that underlies this correlation; however, it is consistent with the intuition that burglaries are prevalent in places where residents have left their homes, e.g., to travel to shopping or service centers. Considering different types of movement destinations, we noticed that transit to Residence and Professional & Other Places are informative in almost all types of crimes.

Table 6.2: Correlation Scores with Assault, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.032	0.000
Distance to Small Business *	-0.007	0.000
Arts and Entertainment	0.023	0.000
College and University	0.029	0.000
Event	-0.002	0.744
Food	0.026	0.000
Night Life Spot	0.011	0.048
Outdoors and Recreation	0.016	0.003
Professional and Other Places	0.035	0.000
Residence	0.024	0.000
Shop and Services	0.026	0.000
Travel and Transport	0.010	0.063

Table 6.3: Correlation Scores with Battery, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.055	0.000
Distance to Small Business *	-0.041	0.000
Arts and Entertainment	0.006	0.240
College and University	0.005	0.346
Event	-0.003	0.522
Food	0.017	0.002
Night Life Spot	0.009	0.093
Outdoors and Recreation	0.013	0.016
Professional and Other Places	0.030	0.000
Residence	0.012	0.027
Shop and Services	0.018	0.001
Travel and Transport	0.012	0.022

Normally, the increase of these types of activities implies more people will on the road and leave their home or workplace. Therefore, more guardians are needed when there are more people are preparing to go back home or going to work when make policy for protecting crimes. Meanwhile, not all crimes could be connected to people's movement pattern. Theft, Deceptive Practice and Assault are the three types of crimes related to all types of people's visiting pattern except visiting to Event.

From Tables 6.2 - Table 6.9, we should also note that most correlation scores are positive. These results are consistent with the hypothesis that mere concentration of individuals, regardless of the venue type, increases risk of crime. The only exception to this observation is Narcotics (Table 6.7) and its negative correlation with transitions to College and University venues. One possible explanation for this is the increased security presence at such venues, which could deter such activity. However, correlation scores for distance to school

Table 6.4: Correlation Scores with Burglary, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.035	0.000
Distance to Small Business *	-0.033	0.000
Arts and Entertainment	0.005	0.318
College and University	-0.005	0.367
Event	-0.003	0.517
Food	0.001	0.813
Night Life Spot	0.005	0.388
Outdoors and Recreation	-0.004	0.472
Professional and Other Places	-0.004	0.418
Residence	0.007	0.170
Shop and Services	0.014	0.012
Travel and Transport	-0.002	0.659

Table 6.5: Correlation Scores with Criminal Damage, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.040	0.000
Distance to Small Business *	-0.033	0.000
Arts and Entertainment	0.018	0.001
College and University	0.009	0.091
Event	0.001	0.822
Food	0.007	0.201
Night Life Spot	0.006	0.231
Outdoors and Recreation	0.002	0.748
Professional and Other Places	0.021	0.000
Residence	0.012	0.068
Shop and Services	0.013	0.000
Travel and Transport	0.007	0.306

and small business are negatively correlated with intensity of crimes. When the distance to each venue type increase, the risk of crimes will be lower. Thus, the distance to some locations like small business, schools and so on are negatively correlated with count of criminal activities. However, where there are more people go to these places, there will be more crimes. The intensity of routine activities are positively correlated with crimes in general. In other words, people's movement to leave or get away from certain locations will increase their chances of facing crimes while stay in schools will be much safer. This finding could support our hypothesis that people's routine activity is correlated with crimes. This finding suggest that the guidance should be allocated to places that near venues like small business. And more guidance are needed when there are more people go to these venues whether it is their daily routine activities or there are some certain events.

Based on our results, we also note that certain venue types exhibited similar correlation scores. For

Table 6.6: Correlation Scores with Deceptive Practice, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.028	0.000
Distance to Small Business *	-0.037	0.000
Arts and Entertainment	0.063	0.000
College and University	0.114	0.000
Event	0.000	0.948
Food	0.048	0.000
Night Life Spot	0.032	0.000
Outdoors and Recreation	0.047	0.000
Professional and Other Places	0.095	0.000
Residence	0.037	0.000
Shop and Services	0.053	0.000
Travel and Transport	0.064	0.000

Table 6.7: Correlation Scores with Narcotics, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.054	0.000
Distance to Small Business *	-0.029	0.000
Arts and Entertainment	0.006	0.261
College and University	-0.012	0.024
Event	-0.004	0.483
Food	0.013	0.017
Night Life Spot	0.006	0.286
Outdoors and Recreation	0.006	0.239
Professional and Other Places	.039	0.000
Residence	0.010	0.068
Shop and Services	0.025	0.000
Travel and Transport	-0.006	0.306

example, Food and Night Light Spot venue types show similar correlations across many crime types. One possible explanation for this is the spatial correlation of venue types: Food and Night Life Spots are often located near each other. These kinds of similar result probably related to some inter correlation between the location of venues. This finding suggest that we could probably investigate the correlation between distance to different venue types or the relation between different types of routine activities to improve the next-place prediction problem and crime prediction.

6.3 Preliminary Classification Solution for Crime Prediction

Using people's routine activity to predict crimes will help the law enforcement to establish policy for preventing crimes and allocating guidance resources in the future. With the same observations, we also consider to

Table 6.8: Correlation Scores with Theft, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.058	0.000
Distance to Small Business *	-0.069	0.000
Arts and Entertainment	0.058	0.000
College and University	0.106	0.000
Event	0.000	0.981
Food	0.066	0.000
Night Life Spot	0.059	0.000
Outdoors and Recreation	0.050	0.000
Professional and Other Places	0.112	0.000
Residence	0.047	0.000
Shop and Services	0.072	0.000
Travel and Transport	0.064	0.000

Table 6.9: Correlation Scores with Other Offense, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.020	0.000
Distance to Small Business *	-0.027	0.000
Arts and Entertainment	0.011	0.044
College and University	0.007	0.176
Event	-0.001	0.830
Food	0.012	0.021
Night Life Spot	0.001	0.875
Outdoors and Recreation	0.006	0.239
Professional and Other Places	0.024	0.000
Residence	0.024	0.000
Shop and Services	0.008	0.155
Travel and Transport	0.008	0.152

identify which point will have crimes considering the concentration of destination venues. We put down label points across the city boundary of Chicago including the points from a grid of evenly spaced points at 2000-meter intervals. The response for each point is labeled with having crimes if there is crime happening in the past one hour or None. The response of this classifier is whether there is any crime in the past one hour in 1000 meters around p. Instead of considering the count of crimes, we build a binary classification model to identify whether the spatial point p is related to any crimes or not. The general form of this classification model is as follows,

$$P(p) = F(Pre(p)) = F(c_1(p), c_2(p), ...c_{10}(p))$$

The probability of crime occurring at a spatial point p will be calculated in some function with features extracted from the occupants to each venue type. Features including $c_1(p), c_2(p), ...c_{10}(p)$ are used to

Table 6.10: Crime Prediction with Next Visiting Venue Count

Result	All Crimes
Accuracy	0.3456
Precision	0.1623
Recall	0.8006
F1	0.1349

characterize spatial point p. We set this function to be support vector machine. The features need to be specified with the result from the output of our Text-Enriched Model for predicting individual's next nearest venue type.

We use the same observations as described in the correlation analysis. Table 6.10 shows the result for our classification model using next nearest venue type features. We get a relatively good recall score which means most of the crimes could be identified. However, considering the data is very skewed, the performance in precision and accuracy get a little lower.

Conclusion and Limitation

7.1 Hypothesis Revisited

Our experimental results support our hypotheses (H1): We find that the textual content of tweets improves next-place prediction compared with baselines that do not consider tweets' textual content. We further find evidence of correlation between predicted next-place concentrations at various venue types and the occurrence of crimes. This evidence support our hypothesis to apply next-place prediction to crime prediction (H2) that crime rates correlate with the intensity of people's movement to each venue type in the same area. However, there are some challenging problems need further investigation. For the next place-prediction problem, we have demonstrated that the textual content is correlated to crimes but we did not investigate further about whether there is any causality between the textual content and routine activities. For the crime prediction, we have showed some evidence to support the Routine Activity Theory which suggests that the crime are positively correlated to the movement to some venue types. This result could benefit for some commercial purpose and prevention of crimes for government. However, we did not build an appropriate solution to implement the result of next-place prediction to crime prediction, so that it is still unknown that how to utilize people's routine activity to predict crimes.

7.2 Limitations

7.2.1 Lack of Ground Truth

The locations and movement trajectories are extracted from Twitter and Foursquare, so we did not have any ground truth to confirm what is the exact place visited by each person. Firstly, the accuracy and reliability of consumer-grade GPS receivers are different in a variety of landscape settings [43], which may expect positional accuracy within approximately 10 m in closed canopies. Thus, the location may not be the correct place for that Twitter user. Secondly, there is no way to identify whether it is the correct place for the undertaken routine activity or not. For example, a user probably post a tweet only 5 meters away from a Food venue, but he just go there for work and have nothing to do with Food venue. We did not know the exact routine activity and place they visit just with the help of check-ins. The only solution to this problem is to do some follow up surveys to each user to confirm their real routine activity or type of places they have visited. To some cases, there is not a straightforward relationship between online content and activities in social media. Researchers [44] have showed that predict real world behavior is much more difficult than extract online activities. Thus, in our cases, we need to do further investigation to identify the relation between online activity and real world behavior to build the ground truth for next-place prediction problem using Twitter. In our work, we assume that the nearest venue type is likely to be the exact type of place visited by each individual. However, if we could not validate the ground truth to identify people's movement pattern in the real world, the ability of using next-place prediction to model routine activity and its other applications will be jeopardized. In the future, We could try to confirm our assumption with the help of Geographical Information Systems in the future to fill the gap in this limitation.

7.2.2 Causality Relationship

As demonstrated in the above sections, using tweets could get significantly improved result for next-place prediction problem and the predicted result is correlated with crime rates. We did not identify any causal link within within the next-place prediction itself and the relation between routine activity and crime rates. For the causality relationship within next-place prediction, it is hard to say how the textual content cause the visiting to some places. Normally, there will not be any direct connection to show such clues. For example, we have found that a user will visit certain types of places in the future through implicit reasons, but we still did not identify why this place will be visited by this user.

Also, even though we have identified some positively correlation between crime rates and occupants visit to each venue type, we did not know what is the casual relation between them. In the future, we are interested

in knowing this causality will improve crime prediction or not in the future.

7.2.3 Representatives of total population

There is a major problem in our methodologies which could only be applied to the city have a large population using Twitter. Also, the people who use Twitter only take a small portion of people within the total number of populations. For example, Twitter users are almost the youth which could not represent the whole population. Researchers [45] has showed that the Twitter population could not demonstrate the demographically characters of the total population. In this case of crime prediction, the populations probably overlap a lot between the potential offenders and victims but may not do well in other applications of next-place prediction problems. Meanwhile, it has been demonstrated that Twitter carries significant predictive power in real world activities [12, 13, 14, 15]. Although Twitter's population did not demonstrate the total population's demographic, we could still utilize Twitter for a variety of applications. Also, the population of offenders or victims could not represent the total populations as well. Therefore, it is useful to figure out the demographics of Twitter's population and its connection with other real world activities which would be helpful to future tasks.

7.2.4 Privacy and Safety Concerns

Our work suggest that it is possible to monitor and predict each individual's movement pattern through social media postings. Although people did not demonstrate any explicit clues about their activities, their routine activities will still be extracted by some statistical methods. This will raise some concern about people's privacy. There has been increasing threats in location-related privacy in geo-social networks [46]. The LBSN let users check-in at certain locations with a geotagged resources which could be accessed by many other users. As we have showed in our work, other users who have access to such content could be potentially to know that user's visiting plans in the future. For example, the offenders could extract information like an individual's home/work location and predict he/she will leave his work/home for which place to be the next visiting place. Researchers [47] has done a simulation to a scenario that the attackers will reveal the identity of a set of LBSN users through their historical check-ins. They have identified user's with certain types of check-in patterns are more likely to be be monitored for a malicious user. Thus, if our methods are applied for extracting individual's routine activity and predict crimes in the future, we should take some actions to protect each user's privacy.

7.2.5 Temporal Effect

Our work did not consider the temporal effect both in next-place prediction problem and the correlation analysis. For the next-place prediction problem, we only considered which venue would be the next visiting place but we did not consider when this venue will be visited. This part of the work need further investigation to build a spatial temporal framework for the next-place prediction problem. For the correlation analysis, our work conclude that people's next visiting occupants are correlated to the happening of crimes in certain areas without further examining the correlation between it with crimes happened in the future.

7.3 Potential Impact

There has been a lot of software has been developed in crime prediction policy making [1][2]. First, we show evidence to support the environmental criminal theory again which need further investigation to understand the relation between crime offenders and routine activity. We have identified that the occupants of movement to each venue type is positively correlated with count of crimes. Thus, it is possible to combine the routine activity with historical criminal density to predict crimes in the future. Second, we have showed that the potential predictive ability in user generated content. This type of content has been applied in areas like president election, sentiment analysis of some news events and so on. We have showed that this type of content could probably be applied in predicting areas related to user's movement in the future.

Future Work

Future work should consider a few major aspects of these problems. It would be interesting to give further consideration to the network of relationships present in the Twitter data. In our experiments, we did not find benefit in using the @-link information for next-place prediction, but this was just a preliminary model and we believe that further investigation might uncover interesting correlations between the venue trajectories of users' friends and the users' themselves. Also, there have been some work to identify different types of @-link which could help us to identify which relations are more likely to be related to users' movement pattern. Moreover, there are some other social relations like following in tweets which could probably provide some networked relation as well. In the venue distance regression setting, one might expect to observe correlations between distances to certain venue types. For example, one might expect to see restaurants near arts &

[1] https://www.predpol.com/
[2] https://www-01.ibm.com/software/analytics/spss/11/na/cpp/

entertainment venues. Thus, the prediction of a user's future distances to such venues should be constrained to take such correlations into account.

Meanwhile, it is interesting to do some deeper text mining to analyze the textual content with its matched locations. Incorporate features extracted from textual content could improve the performance of next-place prediction. However, it is hard to conclude the causality between textual and locations. We could also apply topic model to the documents for each venue type or extract events happening in different places to know better about the routine activity in more detail with the help of social messages.

Regarding crime, we have found preliminary evidence of correlation between predicted next-places and the occurrence of crime; however, we have yet to incorporate these correlations into a full crime prediction model, such as the ones described by [14]. Also, we have only used Twitter and Foursquare to be our data source to model individuals movement pattern in certain areas for predicting crimes. It is necessary to utilize more demographic social statistics to predict and evaluate people's routine activity for crime prediction. Future research focus on crime prediction together with people's routine activity will help the law enforcement build policy for preventing crimes in the future.

Bibliography

[1] Daniel Ashbrook and Thad Starner. Learning significant locations and predicting user movement with gps. In *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on*, pages 101–108. IEEE, 2002.

[2] Fernando Miró. Routine activity theory. *The Encyclopedia of Theoretical Criminology*, 2014.

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

[4] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM*, volume 12, pages 1038–1043. Citeseer, 2012.

[5] Manoranjan Dash, Hai Long Nguyen, Cao Hong, Ghim Eng Yap, Minh Nhut Nguyen, Xiaoli Li, Shonali Priyadarsini Krishnaswamy, James Decraene, Spiros Antonatos, Yue Wang, et al. Home and work place prediction for urban planning using mobile network data. In *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, volume 2, pages 37–42. IEEE, 2014.

[6] Huiji Gao, Jiliang Tang, and Huan Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[7] Defu Lian, Vincent W Zheng, and Xing Xie. Collaborative filtering meets next check-in location prediction. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 231–232. International World Wide Web Conferences Steering Committee, 2013.

[8] Takuya Shinmura, Dandan Zhu, Jun Ota, and Yusuke Fukazawa. Destination prediction considering both tweet contents and location transition hitstory. In *Mobile Computing and Ubiquitous Networking (ICMU), 2014 Seventh International Conference on*, pages 95–96. IEEE, 2014.

[9] Eric Malmi, Trinh Minh Tri Do, and Daniel Gatica-Perez. From foursquare to my square: Learning check-in behavior from multiple sources. In *ICWSM*, 2013.

[10] Jihang Ye, Zhe Zhu, and Hong Cheng. Whats your next move: User activity prediction in location-based social networks. In *Proc. of the SIAM International Conference on Data Mining*. SIAM, 2013.

[11] Jan H Kietzmann, Kristopher Hermkens, Ian P McCarthy, and Bruno S Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.

[12] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[13] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.

[14] Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[15] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[16] Lawrence W Sherman, Patrick R Gartin, and Michael E Buerger. Hot spots of predatory crime: Routine activities and the criminology of place*. *Criminology*, 27(1):27–56, 1989.

[17] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[18] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[19] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[20] Le Hung Tran, Michele Catasta, Lucas Kelsey McDowell, and Karl Aberer. Next place prediction using mobile data. In *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*, number EPFL-CONF-182131, 2012.

[21] Stuart J Barnes and Eusebio Scornavacca. Mobile marketing: the role of permission and acceptance. *International Journal of Mobile Communications*, 2(2):128–139, 2004.

[22] Michael F Goodchild and J Alan Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

[23] Zhongqi Lu, Yin Zhu, Vincent W Zheng, and Qiang Yang. Next place prediction by learning with multiple models.

[24] Trinh Minh Tri Do and Daniel Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 163–172. ACM, 2012.

[25] Trinh Minh Tri Do and Daniel Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, 12:79–91, 2014.

[26] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[27] Dan Jiang, Xiao Guo, Yong Gao, Jiajun Liu, Haoran Li, and Jing Cheng. Locations recommendation based on check-in data from location-based social network. In *Geoinformatics (GeoInformatics), 2014 22nd International Conference on*, pages 1–4. IEEE, 2014.

[28] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[29] Huiji Gao, Jiliang Tang, and Huan Liu. Mobile location prediction in spatio-temporal context. In *Nokia mobile data challenge workshop*. Citeseer, 2012.

[30] Jonathan Chang and Eric Sun. Location 3: How users share and respond to location-based data on social networking sites. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 74–80, 2011.

[31] Jesus Mena. *Investigative data mining for security and criminal detection*. Butterworth-Heinemann, 2003.

[32] Patricia Brantingham and Paul Brantingham. Criminality of place. *European Journal on Criminal Policy and Research*, 3(3):5–26, 1995.

[33] Ronald Victor Gemuseus Clarke and Marcus Felson. *Routine activity and rational choice*, volume 5. Transaction Publishers, 1993.

[34] Patricia L Brantingham and Paul J Brantingham. Notes on the geometry of crime. *Environmental criminology*, 1981:27–54, 1981.

[35] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding hotspots. 2005.

[36] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[37] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[38] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[39] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[40] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[41] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[42] Paul Baumann, Wilhelm Kleiminger, and Silvia Santini. The influence of temporal and spatial features on the performance of next-place prediction algorithms. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 449–458. ACM, 2013.

[43] Michael G Wing, Aaron Eklund, and Loren D Kellogg. Consumer-grade global positioning system (gps) accuracy and reliability. *Journal of forestry*, 103(4):169–173, 2005.

[44] Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, and Kiran Sagoo. Real-world behavior analysis through a social media lens. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 18–26. Springer, 2012.

[45] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

[46] Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, and Christian S Jensen. Location-related privacy in geo-social networks. *Internet Computing, IEEE*, 15(3):20–27, 2011.

[47] Luca Rossi, Matthew J Williams, Christoph Stich, and Mirco Musolesi. Privacy and the city: User identification and location semantics in location-based social networks. *arXiv preprint arXiv:1503.06499*, 2015.