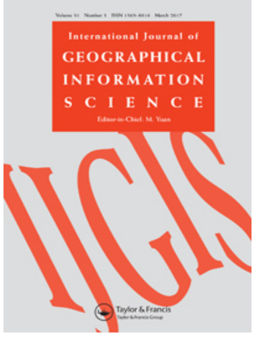


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Mining online footprints to predict user's next location

Qunying Huang

To cite this article: Qunying Huang (2017) Mining online footprints to predict user's next location, International Journal of Geographical Information Science, 31:3, 523-541, DOI: 10.1080/13658816.2016.1209506

To link to this article: https://doi.org/10.1080/13658816.2016.1209506

Published online: 02 Aug 2016.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2017
VOL. 31, NO. 3, 523–541
http://dx.doi.org/10.1080/13658816.2016.1209506

Taylor & Francis
Taylor & Francis Group

Mining online footprints to predict user's next location

Qunying Huang

Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT
Social media applications are widely deployed in mobile platforms equipped with built-in GPS tracking devices, and these devices have led to an unprecedented collection of geolocated data (geo-tags). Geo-tags, along with place names, offer new opportunities to explore the trajectory and mobility patterns of social media users. However, trajectory data captured by social media are sparsely and irregularly spaced and therefore have varying degrees of resolution in both space and time. Previous studies on next location prediction are mostly applicable for detecting the upcoming location of a moving object using dense GPS trajectories where locations are recorded at regular time intervals (e.g., 1 minute). Additionally, point features are commonly used to represent the locations of visits, but using point features cannot capture the variability of human mobility. This article introduces a new methodology to predict an individual's next location based on sparse footprints accumulated over a long time period using social networks, and uses polygons to represent the location corresponding to the physical activity area of individuals. First, the density-based spatial clustering algorithm is employed to discover the most representative activity zones that an individual frequently visits on a daily basis, and a polygon-based region is then derived for each representative activity zone. A sparse mobility Markov chain model considering both the movements and online behaviors of the social media user is trained and used to predict the user's next location. Initial experiments with a group of Washington DC Twitter users demonstrate that the proposed methodology successfully discovers the activity regions and predicts the user's next location with accuracy approaching 78.94%.

ARTICLE HISTORY
Received 3 June 2015
Accepted 1 July 2016

KEYWORDS
Trajectory pattern mining; spatial clustering; spatiotemporal mining; mobility

1. Introduction

Understanding human mobility is important research in social science due to its relevance for a wide range of practical applications, including human behavior and migration patterns and the evolution of epidemics and disease spread (Noulas *et al*. 2012). One of the research topics attracting attention from the scientific community is the prediction of an individual's next location: given a known frequented location for an individual, the goal is to discover the next location most likely to be visited. This objective is accomplished using various data mining approaches that analyze historical and geographical data to discover regular patterns.

CONTACT Qunying Huang ✉ qhuang46@wisc.edu

Traditionally, geospatial data supporting human mobility analysis at the local scale have been gathered using three methods (Huang and Wong 2015): (1) surveys, including travel diaries, that record all locations individuals have visited and corresponding activities (e.g., Chen *et al.* 2011); (2) GPS data from GPS tracking devices (e.g., Geolife dataset; Zheng *et al.* 2010), GeoPKDD datasets (Monreale *et al.* 2009); and (3) phone records containing dates, times, and coordinates of cell phone towers routing the calls (Phithakkitnukoon *et al.* 2010). Surveys or gathering travel diary data are extremely time-consuming, being tedious to recruit and train participants and to process the data (e.g., geocoding place names to derive geo-location coordinates). The GPS data are an alternative but relatively expensive option as it is necessary to equip each user with a GPS device. Additionally and similar to gathering data from surveys, collecting GPS data requires recruiting and training participants. The development of communication techniques makes it easy to access user cell phone records to derive the user's position, but it is difficult to recruit participants because of privacy problems, and software has to be developed and installed. For example, to collect mobility behavior datasets, Gambs *et al.* (2012) distributed smartphones to registered participants with each smartphone equipped with a GPS chip, accelerometer, compass, WiFi, and a Bluetooth interface. The software installed on the smartphones identifies the GPS position and Bluetooth neighborhood of the owner of the smartphone once a minute. A variety of algorithms (Gambs *et al.* 2012) make use of these datasets.

Social media provide another potential source of data describing human daily activity patterns and population dynamics (Huang and Wong 2015, 2016). Social media networks such as Twitter, Google +, and Foursquare have an ever-increasing number of participants. They enable users to share opinions and life experiences in the form of messages. Some messages include publically available, geo-location information of time and location. These social media datasets have rich spatiotemporal (ST) information about people's daily activities and are potentially useful for studying population dynamics and daily activity patterns.

The historical trajectories of a Twitter user are plotted on a Google Map as hotspots at different scales in Figure 1. Displaying the overall trajectories at relatively small scale (Figure 1(a)), it is possible to see some hotspots indicating the daily activity zones (e.g., home, office, parks, coffee shops). Zooming in a region (Figure 1(b)), it is possible to see that this specific user probably travels through Highway I290 from one activity zone to another (e.g., home to office). By further investigating a particular zone area (Figure 1(c)), a general understanding is established about its community or activity zone type (e.g., home or working space) within the geographic context offered by Google Map.

Along with GIS tools, this type of manual interpretation can help analyze and understand an individual's mobility behavior and patterns. However, trying to automatically

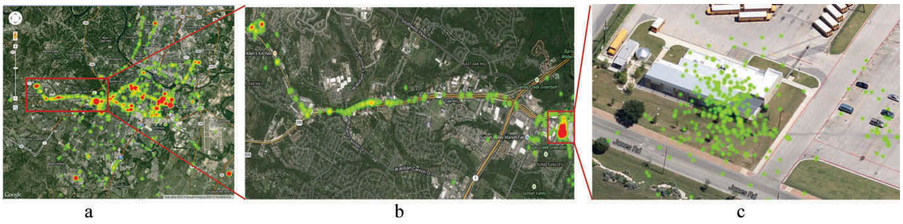


Figure 1. Trajectories of a select Twitter user at different map scales; this user posted 3,611 geotagged tweets between 11 November 2013 and 13 April 2014.

determine the specific type of an activity zone and predict the next location of this individual is challenging. The algorithms or approaches in previous studies are mostly applicable for detecting the next location of a moving object with dense GPS trajectory data where object locations are recorded at regular intervals (e.g., 1 minute) (Yavaş *et al*. 2005, Morzy 2007, Jeung *et al*. 2008) or with data from many moving objects (Alvares *et al*. 2007, Monreale *et al*. 2009) over a short period. With these approaches, meaningful places or point of interests (POIs) are typically derived from GPS trajectories with dense ST points. The POIs are then used to represent the locations a moving object would visit, and trajectory pattern algorithms are developed and used to discover the transition or movement among POIs.

On the other hand, trajectory data using social media are sparse, irregularly spaced, and gathered for a relatively long period, providing data with varying degrees of resolution in space and time (Huang and Wong 2015). Furthermore, if data were gathered over multiple days, a discrepancy might arise because users have different behaviors at the same time but on different days. In fact, individual's commonly visited locations were captured at irregular time intervals, and therefore, it is not possible to systematically know a person's location at a particular time of the day for every day of the week. If points derived from the irregular and sparse trajectories are used to represent an individual's re-visit locations, they may be misleading in showing a higher degree of certainty about the presence of an individual at a specific place and time.

This article investigates the problem of exploring human mobility using social media data and introduces the new concept of region of interests (ROIs) to describe the individual's locations. A sparse mobility Markov chain model (SMMC) is developed to predict an individual's next location for a specific day using historical and current tweets. This article's objectives are the following: (1) select a sample of Twitter users and collect geo-tagged tweets posted in the past several years; (2) detect the most representative or significant places that each user often visited on daily using a spatial clustering algorithm; (3) derive the region boundary of each representative place and use the regions to represent the significant locations; (4) calculate transition matrix among regions and used to build the proposed SMMC model; and (5) evaluate the accuracy of the prediction model using a set of sampled users. This study seeds further research to efficiently and effectively predict an individual next location based on behaviors at an individual level by using social media data.

The article is organized as in six sections. Section 2 presents the related work on the research. In Section 3, the problem and basic definitions related to the proposed prediction model are introduced. The methodology of building the next location prediction model is presented in Section 4, and the experiments and results are reported in Section 5. We conclude in Section 6 with a brief summary and a future research agenda.

2. Related work

2.1. *Next location prediction using GPS trajectories*

Many previous studies (Ashbrook and Starner 2003, Asahara *et al*. 2011, Gambs *et al*. 2012) present the mobility behaviors of an individual as a Markov model or an enhanced Markov model (e.g., mobility Markov chain (MMC); Gambs *et al*. 2012) and predict the next location based on previously visited locations. These models infer possible future visits and corresponding probability for each visit. For example, Ashbrook and Starner (2003) applied the k-means clustering algorithm to automatically cluster GPS data into meaningful locations (e.g., home, work, grocery store), which in turn are used to train a Markov model. Asahara *et al.* (2011) proposed a mixed Markov-chain model that accounts for a pedestrian's personality as an unobservable parameter together with the pedestrian's previous status to predict pedestrian movement. Similar to Ashbrook and Starner (2003), Gambs *et al.* (2011) proposed a mobility model, MMC which was further extended as the *n*-MMC model, to incorporate the *n* previously visited locations in a later study (Gambs *et al.* 2012). Experiments showed that the accuracy and predictability are optimal when $n = 2$, with an accuracy and predictability ranging from 70 to 95%. Based on the work by Gambs *et al.* (2011, 2012), a SMMC model was developed herein to predict user next location using the user's historical online footprint.

There is a strong association among a certain group of users' locations. Through association pattern mining on mobile phone data of 32,579 cell tower locations and 350,000 hours of continuous activity information, Xiong *et al.* (2012) observed that highly confident association rules exist among the locations of users. Therefore, research has focused on predicting an object's next location using data mining approaches to mine frequent trajectories and movement rules from the object's historical locations and to match the trajectory of the object with the mined movement rules for prediction (Morzy 2007). For example, Xiong *et al.* (2012) proposed a collective, behavioral-pattern-based predictor for forecasting an individual's location in the next 6 hours from the locations of other mobile phone users. The results demonstrate the priority in performance over existing predictors.

However, meaningful patterns cannot be extracted from sample points of the moving objects without considering the context of geographic information. Researchers then transform raw into semantic trajectories, which represent the mobility of an object as a sequence of visited places tagged with semantic information (e.g., hotel, church, museum) before applying data mining methods (e.g., Alvares *et al.* 2007, Ying *et al.* 2011). Alvares *et al.* (2007) discovered the most significant parts of trajectories – stops and moves – where stops are the important places for which the duration an object stays exceeds a given threshold and moves are transitions between consecutive stops.

Decision-tree-based mobility models have been developed for next location prediction. For example, Monreale *et al.* (2009) developed and evaluated a decision tree, T-pattern Tree, with a formal training and test process. The tree learns from the trajectory patterns that hold a certain area and is used as a predictor of the next location of a new trajectory by finding the best matching path in the tree.

2.2. *Next location prediction with online digital footprints*

Unlike GPS trajectories, social media online footprints are free and more readily available. Most popular social media platforms offer a public application program interface (API), so researchers can access data from a public stream pool (Huang *et al.* 2014). In the past decade, several attempts have been made to predict the next locations of a single user or multiple users by using various data mining techniques on social media data. For example, Noulas *et al.* (2012) used 35 million user check-ins from Foursquare to define prediction strategies for mining user mobility features for next place prediction based on linear regression and M5 model trees. An M5 model tree includes a linear regression model at each leaf to predict the response of the instances that reach the leaf (Quinlan 1992). Noulas *et al*. (2012) reported that an M5 model tree based on the combination of multiple features offered a higher level of prediction accuracy than linear regression. Scellato *et al*. (2011) developed a link prediction system to estimate the information of places and related user activity based on the data gathered from Gowalla, a location-based social network, for users to check in at 'spots' in their local vicinity.

However, these works predict next location based on collective behaviors from a group of users rather than individual behaviors and therefore ignore the potentially unique ST patterns hidden behind an individual's mobility behaviors and activities. Additionally, current research relies on user 'check-in' data from the location-based social networks, such as Foursquare (Noulas *et al*. 2012). These data may not capture the real trajectories of individuals since many users are unwilling to check in while visiting unimpressive places (Joseph *et al*. 2012). On the other hand, social network sites (e.g., Twitter), can automatically attach geo-tags to messages if the devices have enabled location services. Thus, online footprints collected from such sites may potentially capture more diversity and variability in human mobility. This article develops a new approach to predict next location by considering only an individual' historic movements from Twitter datasets.

2.3. *Spatial clustering*

As one of the most widely used data mining methods, spatial clustering groups similar spatial objects into clusters and gains insight into the characteristics of each cluster (Mennis and Guo 2009). It typically serves as an important preprocessing step for other data mining algorithms or models, such as classification and forecasting models (Miller and Han 2009). Spatial clustering over the sample points of historical trajectories detects primary POIs where an individual regularly visits or a moving object stops for a relatively long duration. In research reported herein, individual's geo-tagged messages are used for next location prediction, and the number of messages within each cluster suggests how often the individual visits each location. Therefore, spatial clustering is applied to learn an individual's significant locations before developing and applying models for next location prediction (Ashbrook and Starner 2003, Zhou *et al*. 2004, Alvares *et al*. 2007, Ying *et al*. 2011, Gambs *et al*. 2012).

The *K*-means (Ashbrook and Starner 2003) is a well-known clustering algorithm used to identify POIs. However, *K*-means algorithm needs to have the number of clusters (*K*) specified in advance. This is challenging since appropriate values for the number of places that an individual frequently visits are not known with certainty. The density-based spatial clustering algorithm (DBSCAN) (Ester *et al*. 1996) discovers clusters of arbitrary shape with noises. It does not require specifying the number of clusters in advance but requires two inputs: the minimum number of points (*MinPts*) forming a cluster and the cluster's radius (*Eps*). These two parameters are less likely to be changed within a particular application. To discover ROIs of social media users in an urban area, *Eps* and *MinPts* can be predetermined, and thus, no input is required per user. In research reported herein, the noises in the sample points are the places an individual visits infrequently, since the focus is locations that an individual visits on a regular daily basis. Thus, DBSCAN is selected to cluster these geo-tagged messages into spatial regions.

3. Problem statement

This section introduces the definitions and concepts for predicting a user's next location.

Definition 1. A *Tweet Set (TwS)* is a sequence of geo-tagged tweets posted by a twitter user over a long time span (e.g., two, three years). The TwS is defined as follows:

$$TwS = \{t_0 = <u_0, l_0, t_0, c_0>, \ldots, t_{N-1} = <u_{N-1}, l_{N-1}, t_{N-1}, c_{N-1}>\},$$

where *t* represents a tweet message, *u* represents a Twitter user name, *l* represents the location (*latitude, longitude*) of the user at the time of tweeting, *c* represents the text content of the tweet, *t* is the local time when the tweet was published, and *N* is the total number of tweets per set.

Definition 2. A *Region Set (RS)* is a collection of disjoint spatial regions, each of which represents a place from which a Twitter user frequently tweets. The RS is defined as follows:

$$RS = \left\{R_0, R_{1,} \ldots, R_{K-1}\right\},$$

where *K* is the total number of clusters derived from individual historical tweet records through spatial clustering. Spatial regions are derived from these clusters, and each is represented as a circumscribed polygon that includes all the tweet points in a cluster.

Definition 3. A *Daily Tweet Set (DTw)* is a sequence of tweets a user posts in one day, and it is a subset of a TwS. It is defined as follows:

$$DTw = <t_0, t_1, \ldots., t_{n-1}>,$$

where *t* is a tweet, and *n* is the total number of tweets per user daily.

Definition 4. *Temporal Segment (TS)* of a region R_i is the duration during which an individual tweets most at the region. For example, an individual might most often tweet from a specific coffee shop between 8 and 9 AM. The TS is defined as follows:

$$TS_i = <T_{\text{start}}, T_{\text{end}}>,$$

where T_{start} and T_{end} are the starting and ending times, respectively, that an individual frequently tweets in the region R_i. This concept helps explore the temporal tweeting and mobility behaviors of an individual. Mining TS (method discussed in Section 4.1) is crucial to exploiting the temporal patterns of a user's mobility.

Definition 5. A *Transition Unit* ($TU_{i,j}$) represents the movement from one region (R_i) to another (R_j) sequentially at one time daily and is represented as follows:

$$TU_{i,j} = <R_i, R_j>$$

From our spatial database, we can retrieve the daily tweet set (DTw), from which all transitions among regions can be extracted (method in Section 4.2). A Markov chain-based model is built for predicting the user's next location by calculating the *TU* probability for each pair of regions as $P\ (TU_{i,j})$, the probability of the user moving from region R_i to R_j. If a user does not move from one region to a particular other region, the probability is 0.

4. Methodology

This section presents the methodology for predicting the user's next location based on social media data. A systematic workflow is designed to efficiently collect, process and mine the data. ST analysis derives useful information from the data and discovers predictive rules. By examining the nature of social media data, which is sparse in both space and time, and capturing the regularity and randomness of the movement behaviors, a SMMC model is developed to forecast the user's next location in the upcoming hours, given a previous visit on a specific day. The workflow for processing the social media data and building the prediction model (Figure 2) consists of six components, from data collecting to model building.

(1) Harvest historical tweets of the selected twitter users through Twitter's data access API. These data are stored and managed in MongoDB, a NoSQL (Not only SQL) database management system (Chodorow 2013).
(2) Retrieve the tweets of a particular user, including spatial and temporal information, from the database as a list of the tweet objects t_i (Section 3).

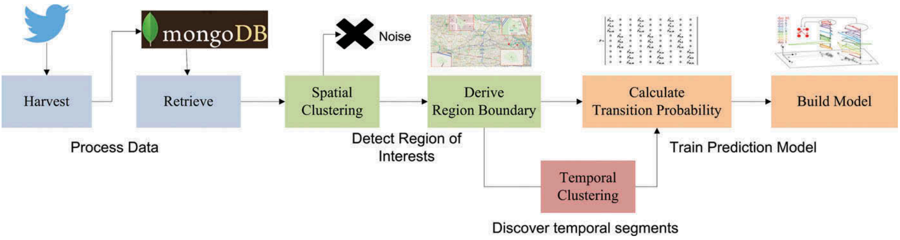


Figure 2. Overall workflow for building the next location prediction model.

(3) Perform a DBSCAN clustering operation on all tweet objects (*TwS*) to identify the most representative clusters, with each cluster including more than a predefined threshold.
(4) Derive the boundary for each representative cluster as a region, with each region (R_i) representing a significant location and assigned a unique ID.
(5) In the temporal clustering process, statistical methods examine the activity's temporal patterns of each region after which each representative region R_i is attributed a temporal segment TS_i.
(6) Discover all possible transitions between pairs of regions for each day and calculate the probability $P(TU_{i,j})$ for each transition to build a transition matrix, with each element represented as a probability.
(7) Develop a SMMC to predict the user's next location visited in a day according to the transition matrix (Step 6) and the region the user previously visited that day.

4.1. *Spatial and temporal clustering*

4.1.1. *Spatial clustering*

Each geo-tagged tweet has a pair of coordinates (latitude and longitude) revealing the specific location where it was posted. Spatial clustering was applied to learn the locations from which an individual tweeted on a frequent basis. As mentioned earlier, the DBSCAN is more appropriate than *K*-means to cluster individual tweet datasets, although it is sensitive to its clustering parameters *Eps* and *MinPts*. Like Zhou *et al.* (2004), the *eps* value was set to 20 meters, approximate to the uncertainty in GPS readings. Ester *et al.* (1996) show that using a *MinPts* value < four may misclassify random points as clusters, whereas a *MinPts* value ≥ four will unlikely produce clusters of varying results. Therefore, a *MinPts* value of four was used.

To reduce the very large number of extracted transitions among places, which in turn results in a very sparse transition matrix (zero value in most the matrix's elements) since many transitions have zero as the transition probability, it is important to remove the places (or clusters) the user only visits irregularly. Since the literature does not provide guidance on differentiate a regular activity from a random place, the value of the minimal number of tweets (*mt*) as a threshold is defined and controls the removal of random activity places. Thus, clusters with geo-tagged tweet numbering < *mt* are considered to noises and were discarded. A lower threshold for the number of geo-tagged tweets per cluster could include more clusters as the representative clusters and therefore potentially provide more detailed transitions to capture a user's daily movements. However, it introduces noise clusters that capture random activities and increases the computing complexity. The impact of different *mt* values to the prediction results are evaluated (Section 5).

4.1.2. *Region of interest detection*

While GPS trajectories have a large volume of dense ST points with very close coordinates to derive POIs, ST points posted through a social network site and collected over a

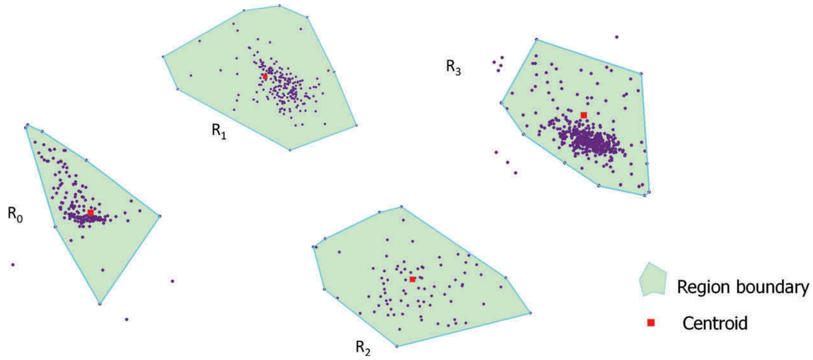


Figure 3. Deriving the circumscribed boundary of each cluster as a region.

long period are sparse and irregular. As discussed earlier, people do not stand and post messages at the same exact coordinates each time they tweet from a region. For example, one day they might tweet from one end of a subway station and another day they might tweet from the other end. We want to make sure that both tweets count as the same location, so using a point feature for each would not capture the individual's activity variety in space and derive the information about the frequency of that individual at that station. Therefore, a polygon feature derived from each cluster is proposed to represent the visited location of the individual geometrically and spatially (Figure 3). Herein, a circumscribed boundary of each cluster including all the points in the cluster represents the shape of a region form where a user may tweet. Simpler features, such circumscribed circles or minimum rectangles, may also be used represent the variability of human daily trajectories.

As a case study, the historical footprints were used from the same Twitter user in Figure 1, located in Austin, Texas. This user had a large number of geo-tagged tweets (3,611 in total), and 27 clusters were derived after performing spatial cluster analysis. To remove places of less interest, the clusters with <30 tweets were removed, resulting in seven qualified clusters and associated regions (Table 1).

4.1.3. *Temporal clustering*

Social media data contain not only spatial information but rich temporal information crucial to mine human movement behaviors over time. Exploring hidden temporal patterns in the historical tweets facilitates our understanding of human mobility and helps predict the user's next location. While social media data, such as tweets, may not reflect the detailed trajectory of a user within a day, they offer select locations of that

Table 1. Spatial and temporal clustering of tweets.

Regions	Number of tweets	Temporal segment (80%)
R_0	438	5–14
R_1	209	6–22
R_2	355	4–23
R_3	98	9–18
R_4	177	0–15
R_5	148	10–23
R_6	476	8–22

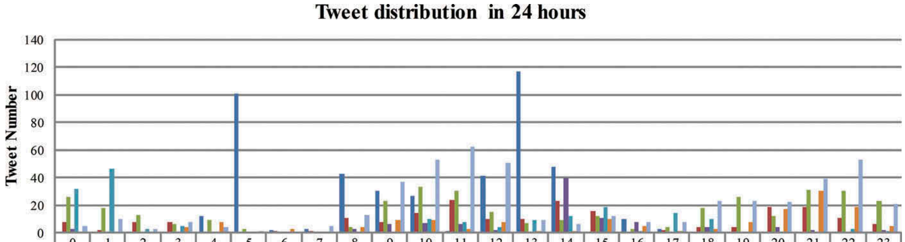
individual over longer periods (several days to years), capturing individual long-term trajectories. Thus, by pooling the location information for multiple days, the longer temporal sample frames compensate for the spatial sparsity of sample points in each day (Huang and Wong 2015). In other words, by combining (or 'aggregating') data from multiple days, the general movements and trajectory pattern of an individual in a daily basis within a period of 24 hours is derived.

The temporal segment or the temporal distribution of tweets within a day at a specific region indicates the period during which an individual tweets or stays in the region, therefore providing clues about the potential location that the individual will visit next. For example, if the total tweeting number is zero at region R_0 between 9 am and 10 am, one can infer that the user usually does not visit this region during that time period. The derived temporal segment of each region is therefore integrated to build our prediction model (Section 4.2).

To derive the temporal segment at each region, the algorithm developed by Huang and Wong (2015) first divides the daily tweets into different temporal window frames using a fixed time interval *I* (e.g., 1 hour) and counts the number of user's tweets for each frame at each region. For example, using 1 hour as the interval (other size intervals are functional as well), the distribution of tweets over 24 1-hour frames is illustrated (Figure 4). For each region, the algorithm starts with the frame with the largest number of tweets. Adjacent temporal frames are incrementally added to form a temporal cluster (*TC*) until the proportion of tweets in the *TC* to the total number of points in the region reaches a pre-defined threshold *l* (e.g., 80%). For region R_0, the 14th frame with the most tweets (117 tweets) is the initial frame to start forming a *TC*. Adjacent frames are then incrementally added to the *TC* until the proportion of points included in the *TC* reaches 80% (assumed *l*) of that in the region. These frames constituting a *TC* form a temporal segment.

4.2. *Sparse mobility Markov chain model*

A discrete-time Markov chain is a stochastic process with the Markov property usually characterized as memoryless: the next stage depends only on the current state and not



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
■ R0	0	0	0	0	12	101	2	3	43	30	27	1	41	117	48	0	10	3	0	0	0	0	0	0
■ R1	8	2	8	8	0	0	1	1	11	8	14	24	10	10	23	16	0	2	4	4	19	19	11	6
■ R2	26	18	13	6	9	3	0	0	4	23	33	30	15	7	9	12	3	4	18	26	12	31	30	23
■ R3	3	0	0	0	0	0	0	0	3	6	7	6	2	0	40	11	8	0	4	0	4	2	0	2
■ R4	32	46	3	5	0	0	0	0	0	0	10	8	4	9	12	19	1	14	10	0	0	0	3	1
■ R5	0	0	0	4	8	0	3	0	4	9	9	3	8	1	1	10	5	1	3	8	17	30	19	5
■ R6	5	10	3	8	4	1	0	5	13	37	53	62	51	9	6	12	8	8	23	23	22	39	53	21

Figure 4. Temporal distribution of footprints aggregated into 24 hours.

on the sequence of events that preceded it (Han *et al*. 2011). The Markov Chain process effectively describes the probabilities of location changes of a moving object over space and time. On the research herein, a SMMC model is used to predict the user's next location using the historical online footprints of the user. The core component of a MMC is to build a transition matrix revealing the probabilities of transition from one place to another. *TU*, which describes a transition from one region to another on one day, is defined as follows:

$$TU_{i,j} = <R_i, R_j>$$

where *i* and *j* are the assigned IDs of regions. In other words, $TU_{i,j}$ represents the user's travel from R_i to R_j. Each transition $TU_{i,j}$ has an associated probability derived for this transition, indicating the probability of moving from region R_i to R_j. Clearly, the higher the probability of a transition, the more frequently the user moves among regions of the transition.

The following algorithm (Algorithm 1) shows the procedure for extracting all possible TUs of the regions that a user would frequently visit and obtaining the probabilities for these transitions. The input of the algorithm is all collections of daily tweets (*DTw*'s), and the output of the algorithm is a transition matrix containing the probability for each transition.

Algorithm 1: Transition matrix

```
Input: All daily tweets (DTw's)
Output: transition matrix containing the probability for each transition
1: for each daily tweet set $DTw_i$ (i = 1 to DTw's.length)
2:          sort the tweet records in the $DTw_i$ by their posting time;
3:          get the start tweet $t_s$ from $DTw_i$ ($t_s = DTw_i$ (0))
4:          get the region ID $r_s$ of $t_s$ ($r_s$ = region_discover ($t_j$))
5:          for each tweet $t_j$ (j = 1 to $DTw_i$.length)
6:               get the region ID $r_j$ of $t_j$ ($r_j$ = region_discover ($t_j$))
7:               if the region ID ($r_j$) is null then
8:                    continue
9:               if the region ID ($r_j$) is not equal to $r_0$ then
10:              increment the count of transition between $Rr_s$ and $Rr_j$ ($TUr_s$, $r_j$)
11:              increment the total number of transitions starting from $Rr_s$
12:              set the value of $r_s$ as $r_i$ ($r_s = r_i$)
13:     end for
14: end for
15: calculate the probability for each transition
16: return the transition matrix with the probability for each transition
```

The algorithm starts by scanning the collection of each day's tweets and then sorts the tweets according to the posting time during that day (line 2). Given the *x* and *y* coordinates of a tweet and the boundaries of all regions, a simple function *region_discover()* is used to discover the ID (r_s) of the region of the first sent tweet and uses it as the starting region (line 4). Subsequently, all subsequent tweets are scanned to check the next region the user visited during that day. The user may send multiple tweets during different time periods in a single region, meaning that during this period the user may leave this region go elsewhere that outside the area not captured by the social media data and then return to the same region. Since it is challenging to check whether this user stayed or moved around using the consecutive tweets posted in the same region, unary transition, which is the transition from one region to itself, is therefore not considered in our work. Accordingly and if a tweet is sent from the same region, this tweet is skipped, and the analysis proceeds to the next tweet. If the tweet originates from a new region, the count of transition between the start region Rr_s and Rr_j (TUr_s, r_j) is incremented, and the value of the control variable (new start region of a transition) is reset as the ID of this region (lines 9–12). The tweets that do not belong to any region are considered as random movement behaviors and discarded (lines 7–8).

The iteration (lines 1–14) is repeated until all *DT*'s are scanned to extract all *TU*s with two sequential regions. For each transition $TU_{i,j}$ of moving R_i to R_j, the probability of $P(TU_{i,j})$ is calculated (line 16). For example, on a specific day (e.g., 3 December 2013), a user posts 8 tweets from 9 to 11 am at region R_0, 10 tweets from 11 am to 14 pm at R_3, 4 tweets from 14 to 20 pm at R_5 and 20 tweets from 20 to 23 pm at R_0 region. In this case, the following three transitions occurred: $\mathrm{TU}_{0,\ 3}$ (R_0–>R_3); $\mathrm{TU}_{3,\ 5}$ (R_3–>R_5); and $\mathrm{TU}_{5,\ 0}$ (R_5–>R_0).

Looping all historical days with tweets available, a variety of *TU*s are produced. If the user has N frequently visited ROIs involved in his daily activities, $N \times N$ numbers of $TU_{i,\ j}$ ($i = 0$ to $N - 1$; $j = 0$ to $N - 1$) are derived in total. Using the probabilities of potential TUs to build the transition matrix of the Markov-based prediction model based on the tweets of the selected user with seven ROIs detected ($N = 7$) is shown (Table 2).

As proposed by Gambs *et al*. (2011) in their MMC model, $P(TU_{i,j})$ can be estimated by simply counting the number of times the user departs a region to another particular region, and dividing this value by the total number of movements leaving this region. Assuming $N\ (R_i)$ is the number of transitions starting from R_i and $N(R_i, R_j)$ is the number of transitions from R_i to R_j, the frequency of the transition between R_i and R_j, $P(TU_{i,j})$ is calculated (Equation 1) as follows:

$$P\left(TU_{i,j}\right) = N\left(R_i, R_j\right)/N(R_i) \tag{1}$$

However, it is argued that the probability of the next region that the user moves to from the current region not only depends on the frequency of the transition but also other factors, including the importance of the next potential region. For example, assuming a user has three frequently visited regions <R_0, R_1, R_2>, and the frequencies of transitions from R_1 to R_0 and R_1 to R_2 are close (~50% for each transition), we need to decide the

Table 2. Transition matrix of the probabilities of TUs.

	R_0	R_1	R_2	R_3	R_4	R_5	R_6
R_0	$P(TU_{0,0})$	$P(TU_{0,1})$	$P(TU_{0,2})$	$P(TU_{0,3})$	$P(TU_{0,4})$	$P(TU_{0,5})$	$P(TU_{0,6})$
R_1	$P(TU_{1,0})$	$P(TU_{1,1})$	$P(TU_{1,2})$	$P(TU_{1,3})$	$P(TU_{1,4})$	$P(TU_{1,5})$	$P(TU_{1,6})$
R_2	$P(TU_{2,0})$	$P(TU_{2,1})$	$P(TU_{2,2})$	$P(TU_{2,3})$	$P(TU_{2,4})$	$P(TU_{2,5})$	$P(TU_{2,6})$
R_3	$P(TU_{3,0})$	$P(TU_{3,1})$	$P(TU_{3,2})$	$P(TU_{3,3})$	$P(TU_{3,4})$	$P(TU_{3,5})$	$P(TU_{3,6})$
R_4	$P(TU_{4,0})$	$P(TU_{4,1})$	$P(TU_{4,2})$	$P(TU_{4,3})$	$P(TU_{4,4})$	$P(TU_{4,5})$	$P(TU_{4,6})$
R_5	$P(TU_{5,0})$	$P(TU_{5,1})$	$P(TU_{5,2})$	$P(TU_{5,3})$	$P(TU_{5,4})$	$P(TU_{5,5})$	$P(TU_{5,6})$
R_6	$P(TU_{6,0})$	$P(TU_{6,1})$	$P(TU_{6,2})$	$P(TU_{6,3})$	$P(TU_{6,4})$	$P(TU_{6,5})$	$P(TU_{6,6})$

next location the user will probably move to starting from R_1. This is challenging since both transitions have similar frequencies. However, if region R_0 is more important than R_2 to the user, the probability that the user would visit R_0 is higher.

One signal to infer the importance of a region is the number of the tweets posted in the region. A reasonable assumption is that people post more tweets from a region they frequent more often. Additionally, how long a user stays in each region is another signal of the region's importance. The longer a user stays in a region, the greater the probability the user visits that region. Therefore, the importance of a region to derive the probability of a transition is calculated as follows (Equation 2):

$$P(TU_{i,j}) = \begin{cases} 0\ if\ N(R_i, R_j) = 0\ or\ i = j \\ \frac{N(R_i,R_j)}{N(R_i)} + wf * \frac{N(R_j)}{N(all)} + wt * \frac{T_{end} - T_{start}}{24}\ if\ (R_i, R_j) > 0 \end{cases} \tag{2}$$

where *N(R$_j$)* and *N(all)* are the total number of tweets posted in region R_j and all regions (including *R$_j$), T$_{start}$* and T_{end} are the starting and ending times, respectively, that an individual remains in a region as derived through a temporal clustering procedure (Section 4.1.3). The ratio of *N(R$_j$)* to *N(all)* is the frequency that the user posts messages in a region. Similar to the regular Markov transition model, the probability of *P(TU$_{i,j}$)* of unary transition (transition from a region to itself) is 0. Additionally, if there is no transition detected between two regions, its transition probability is also 0. This equation uses the tweeting behaviors in terms of the tweeting frequency at a specific region and the duration (or time segment) that an individual stays and tweets in the region to capture the importance of a region. The w_f and w_t are adjustable variables denoting the weights on tweeting frequency and temporal duration. To differentiate this from the TU matrix (Equation 1), the transition matrix from our proposed method is named Sparse TU, and the corresponding model that learns from Sparse TU built from the historical online footprints for next location prediction is named SMMC.

Since tweeting behaviors vary among users, the values of w_f and w_t should differ while building the Sparse TU from the trajectory data of each user. To determine their appropriate values, an interactive procedure is developed. From the training set, 90% of the data are chosen to build the SMMC models with varying w_f and w_t values, and the remaining 10% of the data are used to validate the corresponding models. Initially, both w_f and w_t *are set* to 0, a double loop (iteration) is used to increment their value by 0.1, an associated sparse TU is produced for the prediction model, and a test is run of the model with the 10% testing data until both their values reach 1. For each user, the pair of w_f and w_t values with the highest prediction accuracy is selected to create sparse TU for the next location prediction.

The probabilities for transitions starting from R_1 for TU and sparse TU with w_f and w_t value as 0.1 and 0.0, respectively, are shown in the third and fourth columns (Table 3). The probabilities that the user moves from R_1 to R_0 are 0.3333 and 0.3563 for TU and for sparse TU, respectively. As stated earlier, unary transition, the transition from a region to itself, is not considered, so that the probability of *P(TU$_{i,j}$)* is always equal to 0. The sum of all probabilities of transitions starting from one region is equal to 1 in the TU matrix, which is not necessary in the sparse TU since additional components (significance of each region) are added in Equation (2) to derive the transition probability between two regions. In fact, the derived transition probability can be further normalized to 1.

Table 3. Probabilities for transitions in Markov model (R_1 as starting point).

Transition unit	Transition	Probability of TU	Probability of sparse TU
$TU_{1,0}$	<1, 0>	0.3333	0.3563
$TU_{1,1}$	<1, 1>	0.0000	0.0000
$TU_{1,2}$	<1, 2>	0.2500	0.2687
$TU_{1,3}$	<1, 3>	0.3333	0.3385
$TU_{1,4}$	<1, 4>	0.0.0833	0.0926
$TU_{1,5}$	<1, 5>	0.0000	0.0000
$TU_{1,6}$	<1, 6>	0.0000	0.0000

However, the prediction results do not change as the prediction is based on the ranking of the transition probabilities (Section 4.3), and therefore these probabilities are not normalized.

4.3. *Next location prediction using sparse TU matrix*

Once a sparse TU transition matrix is derived, one easily predicts the upcoming visit location given the user's previous location on that day. Starting with the identification of the region (Rr_s) from which the tweet was posted using *region_discover* (Algorithm 1, line 4), a transition is selected from the Rr_s with maximum probability in the sparse TU matrix. The destination region corresponding to the transition is the next region the user is most likely to visit next.

Based on Equation (2), the $TU_{i,j}$ between all pairs of regions for the selected user are calculated (Table 4 and Figure 5). It is noted that some transitions between two regions have probabilities of 0. For example, P $(TU_{1,\ 5})$ equals 0, indicating that the user did not move to R_1 from R_5. Based on the next location prediction rule which selects the most probable transition from any given starting region (e.g., R_0), the most frequent transitions for the select user is discovered as follows:

$$<R_0, R_6>, <R_1, R_0>, <R_2, R_6>, <R_3, R_6>, <R_4, R_1>, <R_5, R_6>, <R_6, R_2>$$

Starting from R_0, the user's next possible locations include R_1, R_2, R_3, R_4, R_5 and R_6, and since R_6 has the maximum probability (0.50), this is the user's most probable next location. Similarly, we can reasonably predict the next location of the individual given the sparse TU matrix and any other start region.

5. Experiments and results

To validate the prediction model, trajectory data were used from Twitter users in the Washington, DC metro area (DC). The trajectory data were collected using Twitter's

Table 4. Calculated sparse transition matrix for the selected Twitter user.

	R_0	R_1	R_2	R_3	R_4	R_5	R_6
R_0	0.0000	0.1466	0.1034	0.1238	0.0263	0.1773	0.4996
R_1	0.3564	0.0000	0.2687	0.3385	0.0926	0.0000	0.0000
R_2	0.2849	0.0586	0.0000	0.0290	0.0000	0.1506	0.5488
R_3	0.1564	0.2110	0.0000	0.0000	0.0000	0.3411	0.3584
R_4	0.5230	0.5110	0.0000	0.0000	0.0000	0.0000	0.0000
R_5	0.3645	0.0000	0.1406	0.0295	0.0000	0.0000	0.5372
R_6	0.2730	0.0000	0.4215	0.0468	0.0000	0.3133	0.0000

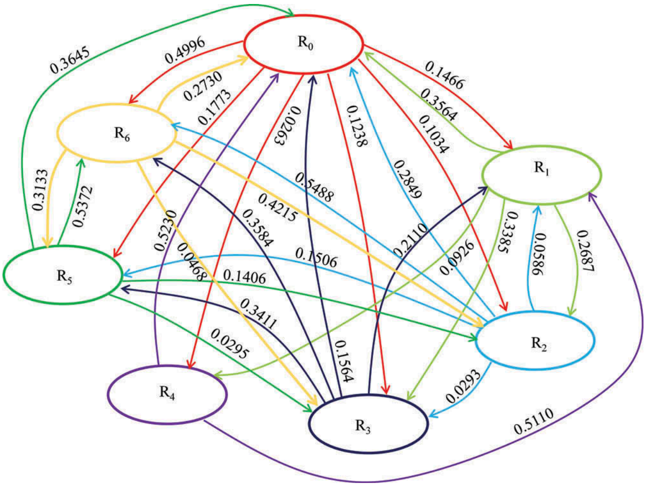


Figure 5. Transition probabilities between all pairs of ROIs.

streaming API with the bounding box of DC for latitude and longitude for the geo-tag filter option. From massive tweets collected between January 2014 and March 2014, more than 3,500 unique users who provided the location 'Washington, DC' in their profiles (retrieved from posted tweets) were identified. Since these users were discovered within the targeted city boundary box and included associated city information in the profile metadata, it is argued that these users live in DC, and their daily mobility and trajectories are valid data source to test the model.

Through a data-driven archiving system developed for harvesting individual-level social media data from multiple social networks (Huang and Xu 2014), the historical tweets of a particular user with the assigned Twitter identification were retrieved. Since tweet harvesting was capped at ≤3,400 tweets per user per harvesting, we continue harvesting more updated tweets for each user after a certain time to collect enough sampling points (geo-tagged tweets). Several user types were unqualified and removed for the study due to three reasons. First, some had geo-tagged tweets <1,000. Most of the selected users with enabled location services included <40% of geo-tagged tweets (Figure 6), resulting in many users are discarded at this step. The second were users with

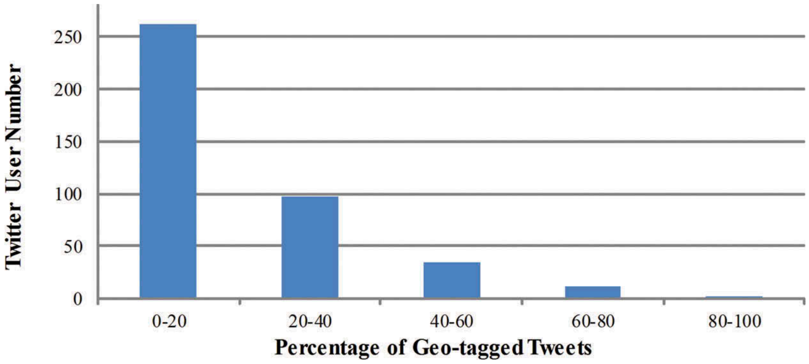


Figure 6. Percentage of geo-tagged tweets by users.

average geo-tagged tweets <6, and it is argued that these users may not have enough daily movements among regions for building the transition matrix. The third were users for whom the geo-tagged tweets were too sparsely located to discover the representative clusters and produce a transition matrix. Of the 3,500 users with DC in their profiles and tweeting geo-tagged tweets from DC and after discarding unqualified users, 52 users were recognized for model validation.

Normally, people have different daily trajectories during weekdays versus weekends. Therefore, only valid geo-tagged tweets posted from these selected users during weekdays were inventoried. The user tweets were separated into two groups, one for training (80%) and the other for testing (20%). The training set was used to build TU matrices and mine historical user trajectories in the model, whereas the latter was selected as the testing sample to evaluate the accuracy of the proposed prediction model for that particular user. The prediction accuracy for each user's model is the ratio of the numbers of correctly predicted regions ($N_{correct}$) to the number of total predictions (N_{total}).

A daily trajectory (*DT*) in the testing sample produces multiple predictions. For example, assuming a user visits five regions in sequence for a specific day represented as <R_1, R_3, R_4, R_2, R_5>, then four predictions are identified. The first region (R_1) is used as the known visited region and from which the next region that the individual will visit is predicted. If correctly predicted as region R_3, the value of $N_{correct}$ is incremented by 1. Subsequently, the next region (R_3) in the *DT* is selected as the known visited region to perform a prediction. This process is repeated until the second to last region (e.g., R_2) in the *DT* is reached.

Figure 7 shows the average prediction accuracy while using different prediction models and different threshold (*mt*) that is used to control the selection of representative regions based on datasets of different users. For each individual, the TU and sparse matrix were constructed using 80% historic datasets and performed validation on the remaining datasets of this individual. As introduced in Section 4.1.1, *mt* determines whether a cluster is a regular activity place or random. A cluster with geo-tagged tweet numbers < *mt* (25, 30 or 35; Figure 7) is regarded as a random place (noise). In general, a higher *mt* results in a higher accuracy (Figure 7). In fact, the prediction based on sparse TU matrix has the highest performance (78.9%) when *mt* equals 35, and this is

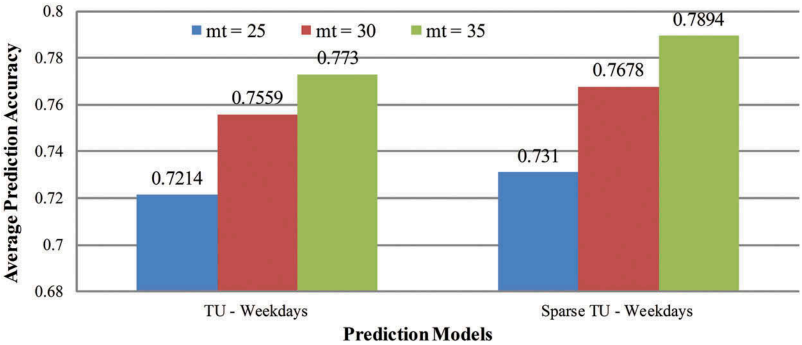


Figure 7. Average accuracy of different prediction models using different threshold (mt) to select representative regions for the selected group of users.

reasonable. By increasing the threshold and reducing the number of noise clusters capturing random activities, fewer clusters are identified as the representative clusters of the user resulting in a simpler transition matrix, which contributes to performance improvement.

Using the sparse TU matrix proposed in the SMMC model in general results in better performance than using only the transition matrix of the MMC model to predict the next location (Figure 7). However, the improvement is not very significant (below 2%). For example, with an *mt* value of 30 and all days' trajectories, accuracies of 75.59% and 76.78% are calculated for the TU and sparse TU models, respectively. After checking the derived values of w_f and w_t (Section 4.2), their values equal zero for most of users' trajectory, meaning the TU model is in general robust enough in predicting the next location prediction. However, it is notable that the prediction for next location is improved for those users when values of w_f and w_t are not calculated as zero based on users' trajectory data. The results indicate that considering users' online tweeting behaviors at different regions, in addition to the frequencies of the transitions that capture users' daily movement behaviors, produce more accurate results for some users.

6. Conclusion

Prediction of a user's next location benefits a variety of applications, including urban planning, transportation, commerce, and homeland security. But an accurate prediction remains a challenge using long-term social media datasets due to the sparseness of the online footprints (geo-tagged messages) and high degree of randomness of user movements. To address this shortcoming, a user's next location prediction framework is proposed from first principles, developed, and tested. Within this framework, polygon-based regions rather than point features are derived from historical online footprints for social media users, and these are used to represent the locations of regular visits. A SMMC model is trained to predict the activity regions in which users are most probable to next appear. Finally, experiments validated the model's accuracy. One Twitter user was used as the case study to illustrate the process of building the SMMC probabilistic model to predict the location the user would visit within hours in a day based on the user's historically visited locations.

While social media data are widely used for mobility studies, these data are not intended for tracking movements and are collected passively (we cannot control when and where data are gathered). As a result, data can only identify the most probable locations from which a user may tweet next. Given a current location, the predicted next location might not be the same as the physical location to which the user transitions. Thus, the predicted next location is more like 'the next tweeting location' while using sparse online trajectory data.

The experimental results using a group of DC Twitter users demonstrate that the proposed SMMC model predicts a user's next location with accuracy approaching 78.94% for weekday data. This model does not include other useful information about an individual's movement behaviors (e.g., user's social network, temporal patterns, local transportation systems, and cultures). These additional sources of information regarding a user's behavior are likely to further improve the reliability and accuracy of the predication model and are the focus of future research.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Qunying Huang http://orcid.org/0000-0003-3499-7294

Reference

Alvares, L.O., *et al.*, 2007. *Towards semantic trajectory knowledge discovery*. Technical Report, October, Belgium: Hasselt University.

Asahara, A., *et al.*, 2011. Pedestrian-movement prediction based on mixed Markov-chain model. *In*: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, 1–4 November 2011 Chicago, IL. ACM, 25–33.

Ashbrook, D. and Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7 (5), 275–286. doi:10.1007/s00779-003-0240-0

Chen, J., *et al.*, 2011. Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography*, 19 (3), 394–404. doi:10.1016/j.jtrangeo.2010.11.002

Chodorow, K., 2013. *MongoDB: the definitive guide*. Sebastopol, CA: O'Reilly Media.

Ester, M., *et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: *The second international conference on knowledge discovery and data mining (KDD*-96), 2–4 August Portland, OR. California: AAAI Press, 226–231.

Gambs, S., Killijian, M.O., and Del Prado Cortez, M.N., 2011. Show me how you move and I will tellyou who you are. *Transactions on Data Privacy*, 4 (2), 103–126.

Gambs, S., Killijian, M.O., and Del Prado Cortez, M.N., 2012. Next place prediction using mobility Markov chains. *In*: *Proceedings of the first workshop on measurement, privacy, and mobility*, 10–13 April Bern, Switzerland. New York, NY: ACM, 3.

Han, J., Pei, J., and Kamber, M., 2011. *Data mining: concepts and techniques*. Burlington, MA: Morgan Kaufmann.

Huang, Q., Cao, G., and Wang, C., 2014. From where do tweets originate? - A GIS approach for user location inference. *In*: *Proceedings of the 7th ACM SIGSPATIAL international workshop on location-based social networks (LBSN '14)*, 4–7 Novermber Dallas, TX. New York, NY: ACM, 1–8.

Huang, Q. and Wong, D., 2015. Modeling and visualizing regular human mobility patterns with uncertainty: an example using twitter data. *Annals of the Association of American Geographers*, 105 (6), 1179–1197. doi:10.1080/00045608.2015.1081120

Huang, Q. and Wong, D., 2016. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographic Information Science*, 30 (9), 1873–1898. doi:10.1080/13658816.2016.1145225

Huang, Q. and Xu, C., 2014. A data-driven framework for archiving and exploring social media data. *Annals of GIS*, 20 (4), 265–277. doi:10.1080/19475683.2014.942697

Jeung, H., *et al.*, 2008. A hybrid prediction model for moving objects. *In*: *Data engineering, 2008. ICDE 2008. IEEE 24th international conference on*. 7–12 April Cancun, Mexico. New York, NY: IEEE, 70–79.

Joseph, K., Tan, C.H., and Carley, K.M., 2012. Beyond local, categories and friends: clustering foursquare users with latent topics. *In*: *Proceedings of the 2012 ACM conference on ubiquitous computing*, 5–8 September Pittsburgh, PA. New York, NY: ACM, 919–926.

Mennis, J. and Guo, D., 2009. Spatial data mining and geographic knowledge discovery - an introduction. *Computers, Environment and Urban Systems*, 33 (6), 403–408. doi:10.1016/j.compenvurbsys.2009.11.001

Miller, H. and Han, J., 2009. *Geographic data mining and knowledge discovery*. Boca Raton, FL: CRC Press.

Monreale, A., *et al*., 2009. Wherenext: a location predictor on trajectory pattern mining. *In*: *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 30 June–1 July Paris, France. New York, NY: ACM, 637–646.

Morzy, M., 2007. Mining frequent trajectories of moving objects for location prediction. *In*: *Proceedings of the 5th international conference on machine learning and data mining in pattern recognition*, 25–27 July Leipzig. Berlin: Springer, 667–680.

Noulas, A., *et al*., 2012. A tale of many cities: universal patterns in human urban mobility. *Plos One*, 7 (5), e37027. doi:10.1371/journal.pone.0037027

Oh, S., 2012. Using an adaptive SEARCH tree to predict user location. *JIPS*, 8 (3), 437–444.

Phithakkitnukoon, S., *et al*., 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. *In*: *Human behavior understanding*, 7 October Vilamoura, Portugal. Berlin: Springer, 14–25.

Quinlan, J.R., 1992. Learning with continuous classes. *In*: *5th Australian joint conference on artificial intelligence*, 16–18 November Hobart. vol. 92. Singapore: World Scientific, 343–348.

Scellato, S., Noulas, A., and Mascolo, C., 2011. Exploiting place features in link prediction on location-based social networks. *In*: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 21–24 August San Diego, CA. New York, NY: ACM, 1046–1054.

Xiong, H., *et al*., 2012. Predicting mobile phone user locations by exploiting collective behavioral patterns. *In*: *Ubiquitous intelligence & computing and 9th international conference on autonomic & trusted computing (UIC/ATC), 2012 9th international conference on*, 4–7 September Fukuoka, Japan. New York, NY: IEEE, 164–171.

Yavaş, G., *et al*., 2005. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54 (2), 121–146. doi:10.1016/j.datak.2004.09.004

Ying, J.J.C., *et al*., 2011. Semantic trajectory mining for location prediction. *In*: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, 1–4 November Chicago, IL. New York, NY: ACM, 34–43.

Zheng, Y., Xie, X., and Ma, W.Y., 2010. GeoLife: a collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*, 33 (2), 32–39.

Zhou, C., *et al*., 2004. Discovering personal gazetteers: an interactive clustering approach. *In*: *Proceedings of the 12th annual ACM international workshop on geographic information systems*, 8–13 November Washington, DC. New York, NY: ACM, 266–273.