Intelligent Automation And Soft Computing, 2019

Vol.25,no. 3,473–486
https://doi.org/10.31209/2018.11007092

Accurate Location Prediction of Social-Users Using mHMM

Ahsan Hussain, Bettahally N. Keshavamurthy, Ravi Prasad K. Jagannath

National Institute of Technology Goa, Ponda, Goa, 403401, India.

ABSTRACT

Prediction space of distinct check-in locations in Location-Based Social Networks is a challenge. In this paper, a thorough analysis of Foursquare Check-ins is done. Based on previous check-in sequences, next location of social-users is accurately predicted using multinomial-Hidden Markov Model (mHMM) with Steady-State probabilities. This information benefits security-agencies in tracking suspects and restaurant-owners to predict their customers' arrivals at different venues on given days. Higher accuracy and Steady-State venue-popularities obtained for location-prediction using the proposed method, outperform various other baseline methods.

KEY WORDS: Foursquare Check-ins, Hidden Markov Model, Location-Based Social Networks, Location Prediction, Temporal Context.

1 INTRODUCTION

LOCATION-Based Social Networks (LBSNs) like, Foursquare, are progressively becoming an important part of a social-user's life, from connecting with friends and sharing images, to discovering new geographical areas and locations. These applications help us to comprehend, investigate, explore, and geographically record the spots we live in. LBSNs permit us to spatially check-in to famous places of a city, updating rich databases that hold computerized engravings of our associations at the same time. The LBSNs Application Programming Interfaces (API's) help to figure out the places where the users are present at any instant of time. This geographic information analysis can uncover the psycho-topography and financial territory of social-users of a city.

LBSN users update and share what they do, where they are and how they feel at a particular time or place. In addition, users uncover when and where they are experiencing a passionate emergency, encountering their very own paradise or damnation, having a good time or a prophetically calamitous occasion. This information, with more conventional government informational indexes, reveals the manner in which the money-related occurrences are linked with these applications. It is a common fact that, more the business importance of an area, more are the number of check-ins from that place. LBSNs can notify a user about the check-in status of their friends to nearby geographical locations with the help of the network itself or by a third-party service provider. User preferences can be explored to enable personalized location-based systems by analyzing the data entered into the LBSNs.

User behavior in LBSNs is highly affected by issues like social relationships and spatio-temporal constraints (Du, Yu, Mei, Wang, Wang & Guo, 2014). (Zhao, He, Zhang, Liu, Zhai, Huang & Liu, 2016; Xiao, Lu & Liu, 2016) analyzed dynamic social-user behavior metrics using trust model and distributed parallel clustering grid respectively in different networks. User preference for various activities is based on the content and context (spatial and temporal) along with the common social interests that help in creation of social communities. The challenge is to comprehensively combine this heterogeneous information in a systematic way, to predict location-based activities of a user. Presence of a user at a certain venue depends on his/her social relationships with the host and other nearby users. The host has a greater role to play in user's preference at a particular venue due to the location and time constraints of activities. So, the host can give better recommendation options for a particular visitor at a place. A privacy mechanism in which user's location data is altered and thus important profits for users are held back, should not be adopted by LBSNs. Users may not like their location information to be hidden from their friends who may be in their vicinity and can join them at an event. The information shared by users through

CONTACT Ahsan Hussain ahsan.hussain@nitgoa.ac.in

LBSNs can be used by security agencies to track a particular user on a given day at some place with the help of a prediction model.

For location-prediction, Hidden Markov Model (HMM), a generative probabilistic model, can be used in which a sequence of observations is generated by a sequence of internal hidden states and cannot be observed directly (Sutton & McCallum, 2006). The inter-state (hidden) transitions resemble first-order Markov chains that are specified by the start probability vector π and a transition probability matrix A. An observation's emission probability on a hidden state gives different distributions with parameter B. This observation sequence can be useful for probabilistic models in case of predicting the hidden state sequences and obtaining the Steady-State probabilities.

In this paper, we analyze LBSN user check-ins and predict the location of LBSN users accurately using multinomial-Hidden Markov Model (mHMM). The digital traces of user check-ins are analyzed to build a user-preference model based on the social relationships and spatio-temporal limits on LBSN users. In addition to user preferences, location of social users is predicted using historical user-venue state sequences. We formulate and parameterize the above elements with the help of user check-ins, into a statistical model using mHMM. It estimates the hidden information in the system by leveraging the known data. In our system, the hidden information which we are calculating, is the location-prediction of LBSN users with the help of previous check-ins. It can be really useful for security agencies in tracking suspects at a particular time based on the LBSN check-in data. At the same time, our model helps the restaurant owners to keep track of their customers' arrivals on particular days and times and also know their food choice pattern well in advance, that can save time and resources. Also, Steady-State probabilities are calculated for all the venues in the city, to formulate the chance of survival for a particular venue in the long-run. We obtain a high accuracy for location-prediction as compared to the other state-of-the-art-methods that formulate the category prediction of venues.

The rest of the paper is organized as follows: the next section includes the description of background and related work in the areas of Foursquare-LBSN and HMM. In section 3, the proposed work for the location prediction after a detailed analysis of LBSNs, is presented. Next, in section 4, the experimental evaluation and thorough result analysis is provided. The complete simulation process is depicted along with an example. Finally, we conclude our paper, along with some future directions, in section 5.

2 BACKGROUND AND RELATED WORK

WE briefly introduce some important existing works done on LBSNs; analysis, categorization and working with a quick grasp of the associated limitations. The focus is on related research involving Foursquare LBSNs and use of HMMs for information prediction, along with a brief look at the Steady-State probabilities using Markov chains.

2.1 Foursquare LBSN

Foursquare, a well-known LBSN, with the main motive of enabling its users to share their geographical-locations (Venues) with friends, has nearly 55 million monthly active users and 8 million average number of daily check-ins on the Swarm app (Foursquare stats, 2017). It has the potential to a review application along with being a social networking service. Check-in is manual and can be done even from a remote location around a city, raising some eye-balls regarding the security of the application. Checking-in at a new place gives rise to a new Foursquare *Venue* and this information is shared in the user-network. Foursquare categorizes its venues with a three-level hierarchical category classification with 9 main groups which are further classified into 291 categories at the second level with incomplete classifications at the third level (Aggarwal, Almeida, & Kumaraguru, 2013; Yang, Zhang, Qu & Cudré-Mauroux, 2016; Noulas, Scellato, Mascolo & Pontil, 2011). Some of the primary venues categorically recognized in Foursquare include Food, Travel-Spots, Great-Outdoors, etc. Users can earn different *badges* as per the number of check-ins at a particular venue. *Mayor* is a special badge given to the user with the most number of check-ins at a restaurant for a period of 2 months. Restaurant owners and other companies offer rewards to Foursquare members and special ones like a free appetizer to its Foursquare mayor. But this can sometimes lead to people cheating to get rewards also. Users can leave positive or negative feedback for a venue visited in the form of publicly visible *Tips*. For putting a tip, however, a user may or may not have to check-in to any place. These tips can help other visitors who plan to check-in to the same place.

Generic and personalized location recommendation have been studied. (Mao, Jiang, Min, Leng, Jin & Yang, 2017) summarized characteristics, design requirements, architecture and surveyed state-of-the-art technologies in mobile SNs. Similarly, (Gasparetti, 2017) studied the existing and future research challenges for location search in LBSNs, which mainly focus on the accurate user preference recommendations, and privacy and serendipity issues. Public opinions for famous venues are used in generic location recommendation, where-in users receive identical suggestions as individual preference is lacking (Cao, Cong & Jensen, 2010). On the other hand, user preference is considered individually in personalized location recommendation systems like matrix factorization (Cheng, Yang, King & Lyu, 2012; Ye, Yin, & Lee, 2010; Ye, Yin, Lee & Lee, 2011) and collaborative filtering (Bao, Zheng & Mokbel, 2012;

Zheng, Zhang, Ma, Xie & Ma, 2011). (Thilakarathna, Seneviratne, Gupta, Kaafar & Seneviratne, 2017) have studied characteristics and evolution of location-based communities based on a social discovery network and geographic proximity.

The venue-location influences users' check-in behaviors (geographical influence), which in-turn helps in making location predictions with new trials to infer users' preferences. The geographical influence can be modelled as a two-dimensional check-in probability density because the probability of a location visited depends on both the location's distance and the intrinsic characteristics (Ye, Yin, Lee & Lee, 2011; Yang, Zhang, Yu & Wang, 2013). (Yang, Zhang, Yu & Wang, 2013) focused on associating location with users and content and scrutinize the potential attacks and protection techniques in GeoSNs for various privacy features including location and identity privacy. The pros and cons of information-disclosure and Privacy in LBSNs have been studied in detail by (Farrelly, 2014); Coppens, Veeckman & Claeys, 2015; Zhao, Lu & Gupta, 2012; Sun, Xie, Liao, Yu & Chang, 2017). Their findings indicate that Foursquare users sought, appreciated and made creative use of the application's geographically relevant place information, enable them to dynamically engage with the place and to create their own meanings. (Litou, Boutsis & Kalogeraki, 2017) have proposed LATITuDE system for efficient dissemination of emergency information, by selecting an appropriate subset of social-users, so that the spread of information is maximized, time constraints are satisfied and costs are considered.

2.2 Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) have been the most effective statistical models in computations and are used to model a lot more complex stochastic processes, as compared to a customary Markov model (Sutton & McCallum, 2006). (Baum, & Eagon, 1967; Baum, Petrie, Soules & Weiss, 1970) have published the fundamental theory related to HMM. HMM is *"a doubly stochastic process with an underlying stochastic process that is not directly observable (it is "hidden") but can be observed only through another stochastic process that produces the sequence of observations"* (Cappé, Moulines & Rydén, 2007). HMM has a set of countable states directed by a set of transition probabilities and an observation sequence can be generated as per an associated probability distribution for every state where only the final result is obtained with the state being invisible to the external observer (Rabiner, 1989).

(Gambs, Killijian & del Prado Cortez, 2012) have extended a mobility model called Mobility Markov Chain (MMC) in order to incorporate the n previous visited locations for next place prediction. They have used the Phonetic, GeoLife and Synthetic datasets with 6, 175 and 1 users, respectively. They incorporated the traces per user, duration of capture, frequency and POI per user as attributes. (Mathew, Raposo & Martins, 2012) present a hybrid method for predicting human mobility as per their characteristics by training HMMs for each cluster and obtained a prediction accuracy of 13.85%, based on the geographical distance between regions over the GeoLife project. (Raghavan, Ver Steeg, Galstyan & Tartakovsky, 2014) developed probabilistic models for temporal activity of social users by incorporating the social network influence as perceived by users based on coupled-HMM.

Continuous Hidden Markov Model (CHMM) was used for developing different recognition models in Intelligent Transportation System that could distinguish regular lane keeping aim from the right and left lane change intentions (Hou, Jin, Niu, Sun & Lu, 2011). (Lane, 1999) studied the deviation from an expected human behavior to detect a malicious social-user by user profiling, based on the posterior likelihood of the model parameters in the HMM. (Li & Li, 2014) introduce a new check-in-based HMM which analyses temporal check-in intervals of users before suggesting locations. To predict the most probable time period, the user will check in next time and also recommend a specific user group for the new entrant. Thus, it can be used to predict the user's check-in location trends with a small but valuable accuracy of 32.54%, 26.39%, and 26.74% for different variations of Gowalla dataset and similar accuracy to some degree for Brightkite datasets. (Ye, Zhu & Cheng, 2013) have used the mixed-HMM to predict the most likely location category, based on user activity at the next step in considerably condensed prediction space. They obtained the best accuracy of 44.35% for category prediction using mixed-HMM. (Gao, Tang & Liu, 2012) have proposed a social-historical model that assesses the role of social correlation in user's check-in behavior in forms of power-law distribution and short-term effect on LBSNs. They have obtained a location-prediction accuracy of 15% - 35% using various models on Foursquare LBSN.

Some basic types of problems that can be solved with HMM (Rabiner, 1989) include: calculating sequence state explaining the observation; probability calculation of the observation sequence for a model; and optimizing the system parameters for obtaining observations. We are trying to solve the latter type of problem in our system using mHMM. For accurate location-prediction, we have implemented mHMM that incorporates the known data of the system and uses it, to estimate the unknown information in the system. The multinomial distribution is the probability distribution of the outcomes from a multinomial experiment, where each trial has a discrete number of independent possible outcomes. We train mHMM by using the spatial and temporal information of users' activities, to further improve the model accuracy.

HMM can be defined as λ = (S, A, B, π) having the following fundamental elements:

(i) The total count of model-states *S*.
(ii) Transitions probability Matrix $A = \{a_{ij}\}$.
(iii) Observation set Matrix $B = \{b_{jk}\}$.
(iv) Initial state values $\pi = \{\pi_i\}$.

2.3 Expectation Maximization (EM) Algorithm

The EM algorithm is an iterative algorithm to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters, where each iteration includes Expectation and Maximization steps (Bilmes, 1998).

1. *Expectation step:*
In this step, using model parameter θ, the probability $P(O|\theta)$ of the given observation sequence *O* is obtained using Eq. (1).

$$P(O|\theta) = \prod_{i=1}^{N} p(o_i|\theta) = L(\theta|O) \quad (1)$$

2. *Maximization step:*
In this step, we try to adjust model parameters θ so that the above probability is maximized.

$$\theta^* = \underset{\theta}{argmax}\, L(\theta|O) \quad (2)$$

The parameters for the new model in terms of old model can be obtained with the following equations for the mixing coefficient $\propto$, mean μ and covariance matrix $\sum$, where $l \in 1,2,..,M$ (Component densities).

$$\propto_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l|o_i, \theta^g) \quad (3)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} o_i p(l|o_i, \theta^g)}{\sum_{i=1}^{N} p(l|o_i, \theta^g)} \quad (4)$$

$$\sum_l^{new} = \frac{\sum_{i=1}^{N} o_i p(l|o_i, \theta^g)(o_i - \mu_l^{new})\,(o_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l|o_i, \theta^g)} \quad (5)$$

Until the difference between the new model and the previous model parameters is less than a certain threshold, the iterations are continued. The above equations are used for both steps of EM simultaneously.

2.4 Viterbi Algorithm

Viterbi algorithm is a recursive programming algorithm that determines the most probable sequence of underlying hidden states from an observation sequence that might have generated it (Forney Jr, 2005). It finds maximum overall state-sequences possible with the help of HMM parameters. The maximum probability of reaching a particular intermediate state is defined by the partial probability δ, ($\delta_t(i)$ being the maximum probability of all sequences ending at state *i* at time *t*). Thus, the best state-sequence can be obtained by the following steps:

1. *Initialization step:*
$$\delta_1(i) = \pi(i) b_i(o_1), \qquad 1 \le i \le N \quad (6)$$

Where $b_i(o_1)$ is the probability of emitting response time o_1 in state *i*.

2. *Recursion step:*
$$\delta_t(i) = max_j\big(\delta_{t-1}(j)\big) * a_{ji} * b_i(o_t) \quad (7)$$

The final step is to determine the arrival of the state *i* at time *t* optimally after calculating the system probability for the previous state at time *t* – 1. A back pointer ∅ points to the previous state that optimally incites the current state in the termination step.

3. *Termination step:*
$$\emptyset_t(i) = argmax_j\big(\delta_{t-1}(j)\big) * a_{ji} \quad (8)$$

Where *argmax* operator selects the index *j* which maximizes the bracketed expression.

2.5 Steady-State Probability

A system is said to be in a Steady-State, if the variables used to define the system-behavior, do not change with respect to time. Markov chain model can be effectively used for obtaining the Steady-State behavior of people visiting a particular venue over a longer period of time. The *n*-step transition probabilities $r_{ij}\,(n)$, can be converged to steady-state values, independent of their initial states. The rows of this limiting matrix contain the probabilities of being in the various states as time gets large. These probabilities are called Steady State (SS) probabilities (Bertsekas & Tsitsiklis, 2002). Thus, the SS probability of state *j*, for large *n*, can be interpreted as

$$\pi_j \approx P(X_n = j) \quad (9)$$

A Markov chain with recurrent aperiodic classes having *j* states with π_j SS probabilities has the following properties.

$$\pi_j = \sum_{k=1}^{m} \pi_k \,.\, P_{kj} \quad (10)$$

$$\sum_{k=1}^{m} \pi_k = 1 \quad (11)$$

Eq. (10) is called the *Balance* equation and Eq. (11) is called the *Normalization* equation. The unique SS solution to the system is given by the quantity π_j, where $j = 1, \ldots, m$. Assuming $\pi_j > 0$, for all recurrent states *j* and $1 \le (i,j) \le m$, the *transition-probability* matrix is :

$$P = [p_{ij}] \tag{12}$$

Using Eq. (12), the SS probabilities for different venues can be obtained that gives the measure of the popularity of various venues surviving at a place.

3 PROPOSED WORK

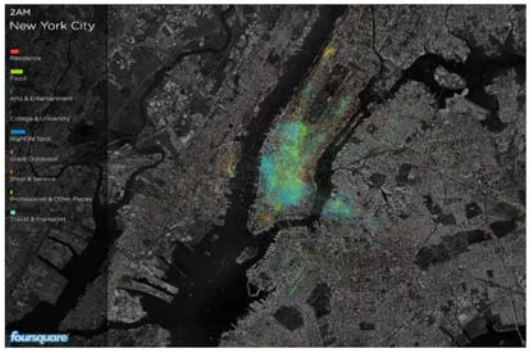
IN the proposed work, we thoroughly analyze the Foursquare check-in patterns in two major world cities, New York and Tokyo in section 3.1, with extensive LBSN user count. It helps in formalizing the preferences of various social-users at different times of a day. Next, the temporal context of users is studied that gives the daily and weekly periodical patterns of their visits to different venues. Finally, we formulate our model using mHMM for location prediction in LBSNs. Our model predicts the exact venue a particular user visits. In addition, we also predict the customers visiting a particular restaurant/venue on a given day of the week. Steady-State probabilities are obtained for depicting the chance of survival of a particular restaurant or venue in the future.

3.1 Analyzing Foursquare check-ins expressing city pulse

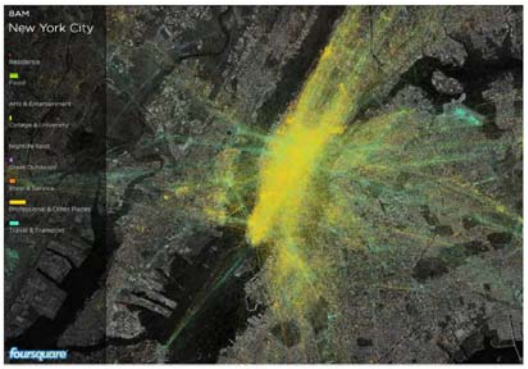
Every day, millions of people check-in at Foursquare. The check-ins data for one year taken from New York and Tokyo cities is plotted on a map29 (Foursquare check-ins pulse, 2017). Interesting visualizations of ebbs and flows of the cities are generated by the start-up data. Single check-ins are represented by individual dots, while the sequential check-ins are linked by straight lines. The condensed check-ins representation shows the city appearance on a usual day. It specifies where people are present at any moment of the day and how, when, and why they're going there. LBSNs users show a more pronounced range of check-ins activity near more commercial places in cities. In the following figures, we depict the snapshots of check-ins done by New York and Tokyo Foursquare users at various times of the day that help in the analysis of the check-in patterns.

3.1.1 Check-in pattern in New York City

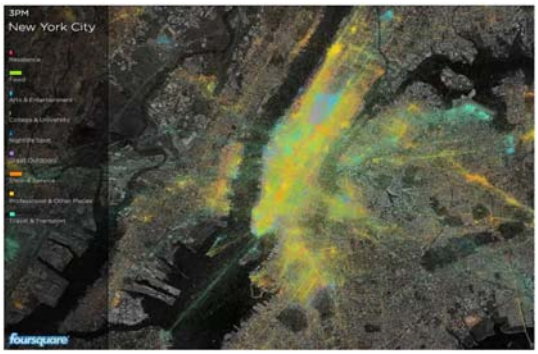
We show instances of check-ins at four different times of a day in New York City in figure 1. Figure 1(a) shows that at 2 AM, the highest number of check-ins are done at *nightlife spots* (Blue), while most of the other places remain silent. However, in figure 1(b), the image captured at 8 AM, depicts a completely different view, where more people get active at this time and check-in at *Professional and other places* (Yellow). As shown in figure 1(c) at 3 PM, *Shop and Services* (Orange) have the most number of check-ins while *General Food places* (Green) are seen abundantly at 8 PM at night in figure 1(d). It shows that once the day sets in, users prefer to go at *Professional Food places* and as the day progresses show an increased presence in outdoor places, with least amount of activity during late nights.



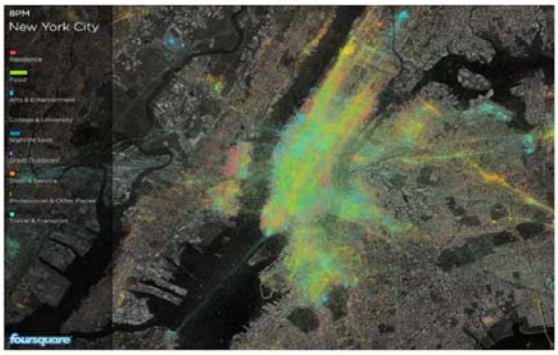
(a) 2 AM



(b) 8 AM



(c) 3 PM

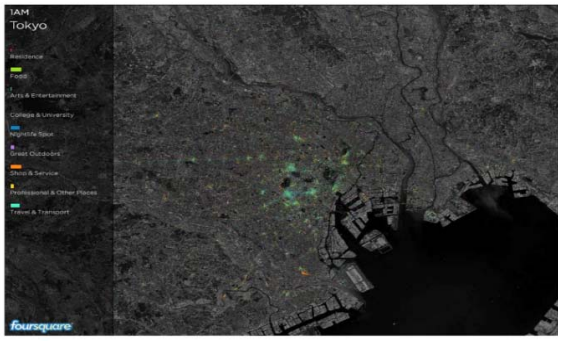


(d) 8 PM

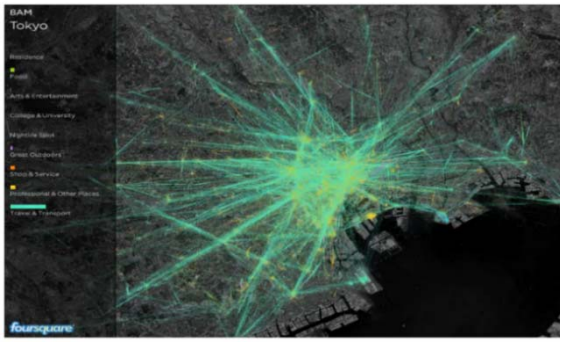
Figure 1. **Check-in patterns in New York City.**

3.1.2 Check-in pattern in Tokyo City

In the following four visualizations of check-ins at Tokyo City (figure 2), it can be clearly seen that most of the check-ins during the whole day occur at venues like *Travel and Transport places* (Green) followed by *Food places* (Indigo). Major activity pattern at the initial part of the day shows great rush at the Travel and Transport places and there is slow but steady increase in the user-activity at other Food places as the day progresses.



(a) 1 AM



(b) 8 AM

(c) 3 PM

(d) 8 PM

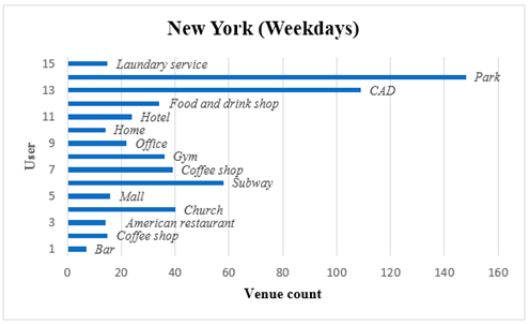
Figure 2. Check-in patterns in Tokyo City.

3.2 Temporal context

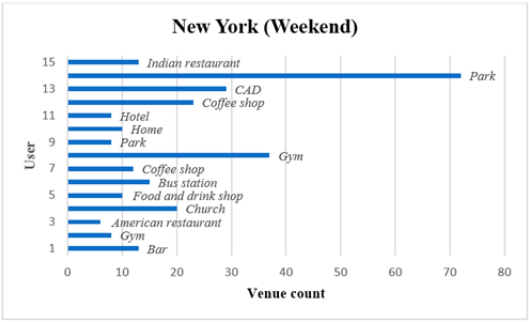
The check-in patterns studied in previous section showed the time-based daily visiting frequency of LBSN users. The temporal context depicts the periodical patterns of user check-ins, where they can show strong daily and weekly periodical affinity for various venues that can help in location prediction. The locations or venues a social-user visits in day-to-day life, generally display a cyclic phenomenon over days of the week. The user preference to different venues can be calculated using the day of the week factor $P_u\ (v)$ in the following Eq. (13):

$$P_u\ (v, v_i) = \begin{cases} 1, d(t_v) = d(t_{vi}) \\ 0, d(t_v) \neq d(t_{vi}) \end{cases} \quad (13)$$

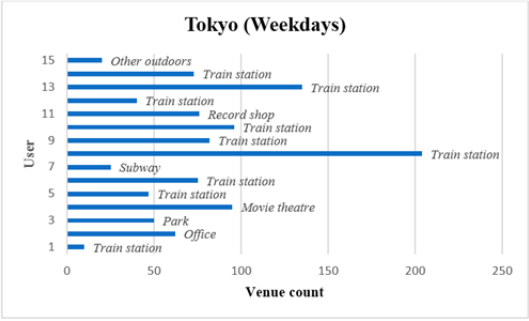
where $d(t_v)$ represents the day when the user visited venue v in the week, and $d(t_{vi}) \in$ {Sunday, Monday, …., Saturday}. Figure 3 shows the venues with the highest number of visits during weekdays and weekends for first 15 users in New York and Tokyo datasets. It depicts the difference between the patterns of visits. For example, at New York, *user-1* visits *bar* more often over the weekend as compared to weekdays. Similarly, *user-2* prefers *Coffee Shop* during weekdays that can be located near his work-place and goes to the *gym* more on weekends. In case of Tokyo, *user-2* visits *office* more in the weekdays and goes to *train-station* on weekends. For *user-14* the scenario is almost reverse, showing more visits to the train-station on weekdays and electronics-store at the weekend.



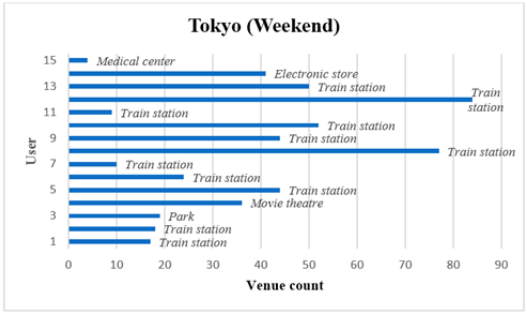
(a)



(b)



(c)



(d)

Figure 3. Temporal Context in Check-in patterns in New York and Tokyo Cities over Weekdays and Weekend.

3.3 mHMM for user location prediction in LBSNs

The proposed model predicts the location of LBSN users with the help of their sequence of visits on previous occasions to particular places. Most of the existing work deals with clustering the venues based on their type or specialty and then predicting the category where the user can visit (Ye, Zhu & Cheng, 2013; Gao, Tang & Liu, 2012). We, in our proposed approach, try to predict the exact location of a user and not the category, as proposed in previous works and obtain a much better prediction accuracy. It makes easier to predict, where a user can visit in an unambiguous way using the proposed model. It is important because in most of the cases of Foursquare venues, multiple venues can exist under same food-category. Therefore, the exact location of a suspect at a particular time, can help the security agencies to track the person efficiently. In addition, by inter-changing the parameters of our model, it can be used by the restaurant owners to know the arrival of their customers on a particular day and time. They can know their customer's food choice and patterns to prepare themselves, well in advance, which can save time and resources. Our model also obtains the SS probabilities that define the survival possibility of a venue over a longer time in future. To the best of our knowledge, our work is the first in the domain of location prediction for users on venue basis of LBSNs.

Our model can be stated by following four parameters:

- ***R:*** set of LBSN restaurants or venues.
- ***A:*** matrix of transition probability. $\{a_{ij}\}$ represents the probability of going from venue *(i)* to venue *(j)*.
- ***B:*** matrix of observation-sequences. $\{b_{jk}\}$ is the observation that a particular user visits a particular place on a given day of the week (this is the unknown parameter to be obtained).
- ***π:*** initial state of the system. We assume constant initial values to the venue-relationships.

We obtain the respective probabilities using the mHMM for our system in the following manner.

(i) N is the Restaurants' (states) count. We designate the set of Restaurants

$$R = \{R_1, R_2, ..., R_n\} \tag{14}$$

where R_i, i = 1, 2,..., N, is a distinct Restaurant.

(ii) M is the number of Days of Week (observation symbols) for every *Restaurant*. We denote the set of symbols

$$V = \{V_1, V_2, \ldots V_M\} \tag{15}$$

where V_i, i = {S, M, T, W, T, F, S}, is an observation symbol (Day of Week).

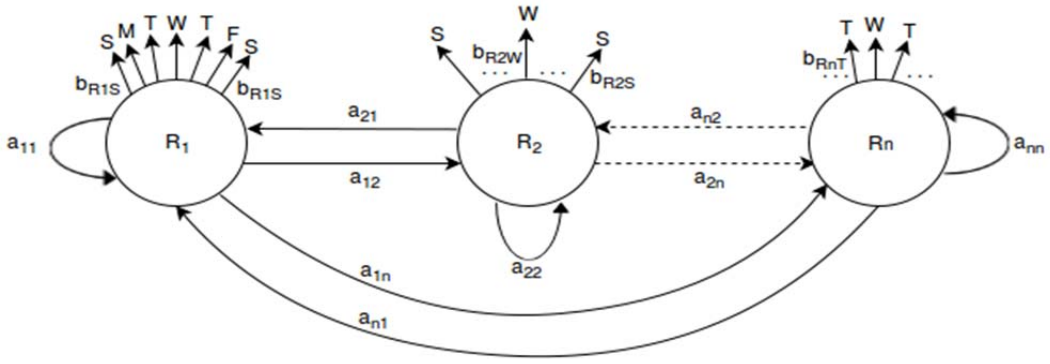


Figure 4. **Proposed mHMM for location-prediction.**

(iii) The transition probability matrix of restaurants

$$A = \{a_{ij}\} \tag{16}$$

where $a_{ij} = P(q_{t+1} = R_j \mid q_t = R_i)$, $1 \le i, j \le N$; $t = 1, 2, \ldots$; for all i, j and $\sum_{i,j=1}^{N} a_{ij} = 1$ and q_t indicates the restaurant (state) at time instant t.

(iv) The probability matrix of observation symbol

$$B = \{b_{jk}\} \tag{17}$$

where $b_{jk} = P(V_k \mid R_j)$, $1 \le j, k \le N, M$ and $\sum_{k=1}^{M} b_{jk}=1$, $1 \le j \le N$.

(v) The initial state probability vector

$$\pi = \{\pi_i\} \tag{18}$$

where $\pi i = P(q_1 = R_i)$, $1 \le i \le N$ and $\sum_{i=1}^{N} \pi_i = 1$.

(vi) The observation sequence

$$O = O_1, O_2, \ldots, O_L \tag{19}$$

where each observation O_t is one of the days of week from V and L is total count.

Algorithm *mHMM*:

Input: Foursquare dataset.

1. Calculate *initial probabilities* of all venues using their membership values.
2. Assign the constant values obtained for transition probabilities to all the venues.
3. Calculate *emission probabilities* as the actual transition count from all the venues (i.e., probability of each venue getting a particular user on a particular day).
4. Input these probability values to the *mHMM* for obtaining the final probabilities of the hidden states (Venues).
5. For all the user-day/venue-day pairs, predict the hidden state sequence based on the actual state sequence.

Output: Particular venue for input user-day pairs.

LBSN user generally follows a trend in the places he visits, as given in detail by (Noulas, Scellato, Mascolo & Pontil, 2011). On weekdays, he will most probably visit the restaurants near his work place as compared to the places visited on weekends. We consider the sequence of places visited by a particular user on a particular day of a week. To map the user locations detection in terms of mHMM, we consider the 7 days of a week as the observation symbols in our model. We quantize the observed values into M days, forming the observation symbols $V_1, V_2, \ldots, V_M$, where V_i, i = {S, M, T, W, T, F, S}, making $M = 7$. Social users on Foursquare give their check-in information at a particular venue. The sequence of the states is hidden from the observer as he is only able to see the final outcome (place visited). The set of all possible types of restaurants visited, forms the set of hidden states in the mHMM.

The next step after deciding the symbol and state representations, is to determine the probability matrices *A, B,* and π. We use *Viterbi* algorithm and *Multinomial-HMM* to determine these three model parameters in training phase of our proposed work.

As shown in figure 4, $R_1, R_2, \ldots, R_n$ denote the various restaurants (hidden states). The transition probabilities are given by a_{ij}, where $(i, j) = 1, 2, \ldots, N$. The transition a_{11} denotes the probability of a user visiting the same venue (R_1). Similarly, a_{12} denotes the *transition probability* of a user in moving from restaurant R_1 to R_2. V = {S, M, T, W, T, F, S} represents the observation symbols from each hidden state. The emission probabilities are given by b_{jk}, where $1 \le j, k \le N, M$. The first emission b_{R1S} represents the emission probability of restaurant R_1 on Sunday and b_{R2W} represents the emission probability of restaurant R_2 on Wednesday. We consider the special case of fully-connected mHMM, where-in, the reachability of every state is at a single hop from every other state in the model.

In a similar way, we reverse the inputs to the model for obtaining the prediction information for the restaurant owners about the user-visits. We formed the Venue-day pairs and these input parameters are given to the mHMM, as in the previous case to obtain the respective transition and emission probabilities. This time, we obtain the output as the predicted set of users visiting a particular venue on a particular day of a week. This can help the restaurant-owners to prepare themselves for the upcoming guests and customers,

based on their number and the type of food-choices they make.

4 EXPERIMENTAL EVALUATION AND RESULTS

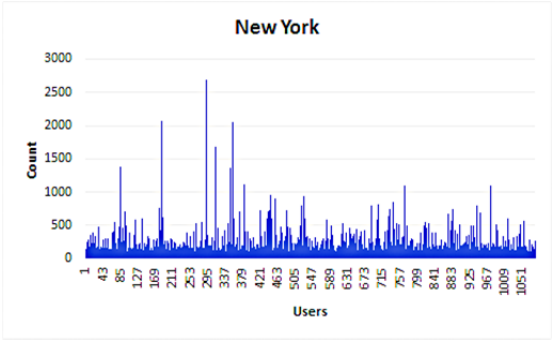
IN the following section, we give the detailed experimental procedure and result analysis for our proposed work.

4.1 Dataset description

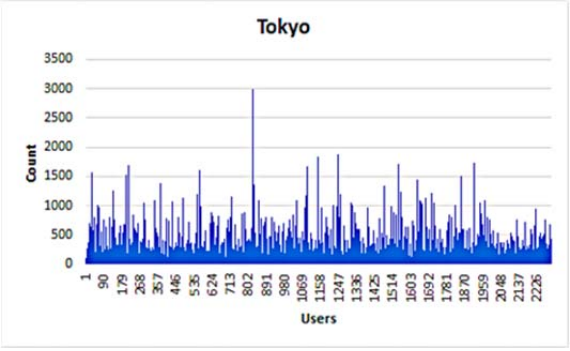
Foursquare check-ins data for 4 months ranging from 24 October 2011 to 20 February 2012 is used for the experimentation purpose (Yang, Zhang, Zheng & Yu, 2015). Noise and invalid data are filtered during the preprocessing. Personal check-in information is not available publicly and can only be accessed from one's own social circle. Only active users with at least one check-in per week are considered and all the check-ins from the "sudden-move" users are eliminated. The category information of some Foursquare venues is unavailable as these cannot be resolved by Foursquare venue API.

In the dataset, we represent a check-in by a quadruple *user-time-location-activity*, where users are uniquely identified by the IDs. User check-ins serve as an indicator of user activities in LBSNs. An activity is represented by venues corresponding to the check-in with the exact geographical coordinates and time. The spatial and temporal dimensions of check-in data are discretized according to the aforementioned LBSN scenario. After preprocessing, the New York City dataset includes 1083 users and 227428 check-ins performed over 251 venues. Similarly, 2293 users registered 573703 check-ins over 247 venues in Tokyo City dataset.

We give the graphical representation of the user check-ins count over the given time period for the two cities in figure 5. It shows the total number of visits a particular user makes during the given duration. For example in New York, *user-293* has the highest number of check-ins (2697) and *user-18* the lowest check-ins (100), with an average of 210 check-ins. Similarly in Tokyo, *user-822* has the highest number of check-ins (2991) and *user-43* the lowest check-ins (100), with an average of 297 check-ins over the given time period.



(a)



(b)

Figure 5. User check-ins count for New York and Tokyo cities.

4.2 Experimental setup (Observation symbol generation)

For each social-user, we train and maintain mHMM and find the respective observation symbols, after pre-processing. All the user-ids from the New York (1083) and Tokyo (2293) cities' datasets were used for our experimentation. 7 individual runs were made for each day of the week for all the 3376 user-ids in both datasets. We separately calculate the transition and emission probabilities for the social users to obtain the accuracy in location prediction. In the following steps, we show sample results obtained using our proposed model for *user-1214* from Tokyo City dataset with *127* check-in entries in *33* different venues and show the sample mHMM output probabilities in 5 venues for the same user in table 1.

Table 1. Sample mHMM probabilities in 5 venues for user-1214

Venue	Entries count	SP	TP	EP
Food & Drink Shop	0,1,0,0,0,0,0	0.0078	0.030303	0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0
Fast Food Restaurant	0,1,0,0,1,0,0	0.0159	0.030303	0.0, 0.5, 0.0, 0.0, 0.5, 0.0, 0.0
Japanese Restaurant	3,2,4,1,1,8,0	0.1508	0.030303	0.158, 0.105, 0.21, 0.053, 0.053, 0.42, 0.0
Shrine	0,0,0,0,0,1,0	0.0078	0.030303	0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0
Airport	1,0,2,0,3,0,0	0.0476	0.030303	0.166, 0.0, 0.33, 0.0, 0.5, 0.0, 0.0

1. *Combinations:* It is the list of the user-day pairs for all user-ids in the datasets. {['1214', 'Wed'], ['1214', 'Sun'], ['1214', 'Thu'], ['1214', 'Mon'], ['1214', 'Fri'], ['1214', 'Tue'], ['1214', 'Sat']}

2. *Entries:* It contains the count of visits by a user to a particular venue on each day of the week. As an example, the user visits the *Japanese Restaurant* 8 times on Tuesdays and *Airport* thrice on Fridays.

Final Probabilities using mHMM

3. *Start Probability (SP):* It is calculated by the individual membership value for each restaurant given all entries of the user. It is given by following Eq. (20).

$$SP = \frac{Individual\ venue\ count}{Total\ venue\ count} \quad (20)$$

4. *Transition Probability (TP):* This is the probability of a user switching between venues, i.e., probability of user going to a restaurant after visiting a particular restaurant or venue (Eq. (21)). The mHMM is built for each location with transitions to every other location. *TP* is set to zero, if the user did not travel between two venues. If the model is trained on the states of link l and its neighbor links in time interval t, the state of link l at time $t+1$ is independent from all other current link states, all past link states, and all past observations.

$$TP = \frac{Trans.from\ venue_i\ to\ venue_j}{Total\ trans.from\ venue_i} \quad (21)$$

5. *Emission Probability (EP):* This is the probability of each venue getting a particular user on each day of the week. It is calculated by the following Eq. (22).

$$EP = \frac{Entry\_value}{Total\_entry\_sum} \quad (22)$$

4.3 Sample output results

The results obtained are in the form of venue-names associated with a particular user for a particular day. We provide the user-id of the social-user, whose location we want to predict along with the day, to get the predicted place of his visit on that given day. We give some sample outputs as follows, where *Bob* is the enquirer (input) and *Alice* is the responder (output).

Bob says: **Wed**
Alice hears: ***Ramen / Noodle House.***
Bob says: **Sun**
Alice hears: ***Japanese Restaurant.***
Bob says: **Fri**
Alice hears: ***Train Station.***

4.4 Result analysis and discussion

We conducted the experimentation on real world datasets for New York and Tokyo cities. We show the final accuracy comparison of results obtained using mHMM for different fractions of training dataset in figure 6. Further, in table 2, we give the minimum, maximum and the average of accuracy results obtained for the various partitions of the datasets. The accuracy varies between 29% to 41% in New York City, with an average of 35% location prediction accuracy. For Tokyo dataset, we obtain a much better accuracy range between 45% to 61%, with the average being 55% accurate location prediction. These results are much better as compared to the other contemporary works in the field of location prediction for LBSNs.

Table 2. Overall Accuracy of venue-prediction.

City	Min.	Max.	Avg.
New York	0.29	0.41	0.35
Tokyo	0.45	0.61	0.55

Table 3. Overall Accuracy of user-prediction.

City	Min.	Max.	Avg.
New York	0.194	0.31	0.257
Tokyo	0.193	0.29	0.258

In figure 7, we show the output results obtained for predicting users visiting a particular restaurant or venue on a given day using mHMM for varying training sets. Table 3 gives the minimum, maximum and the average of accuracy results. On analyzing the results, we understand that user-prediction is a difficult task as compared to venue-prediction. It is because of more number of users than the number of venues due to which the probability of accuracy prediction suffers. Nevertheless, this information is important for the restaurant owners to plan the food-preference patterns for particular customers visiting their places on particular days in advance. This gives a completely different view-point for using our model that can be beneficial in both ways. It can predict the location of a user and by interchanging the parameters, it can be beneficial for restaurant-owners to know the details of their customers visiting on any given day.

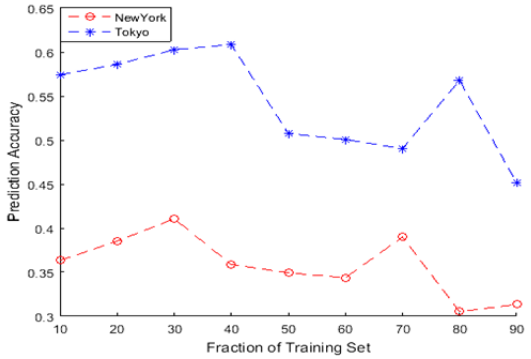


Figure 6. Accuracy of venue-prediction using mHMM.

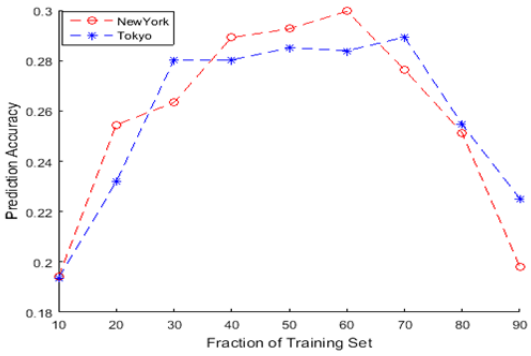
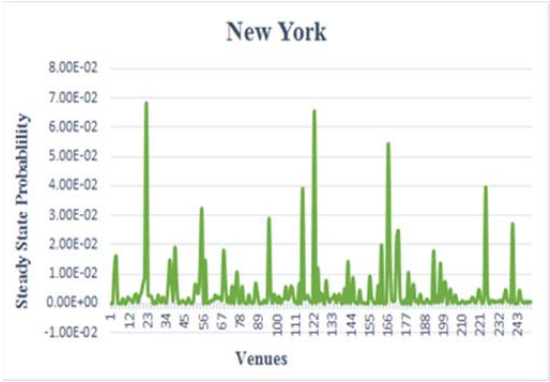
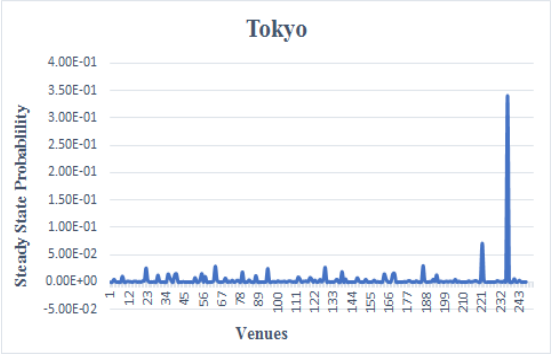


Figure 7. Accuracy of user-prediction using mHMM.

Figure 8 shows the SS probabilities obtained for all venues present in New York and Tokyo City Foursquare datasets. We list SS probabilities of all venues that predict their chance of survival in the future. Out of a total of 251 venues in New York, it can be observed that the best SS probability obtained is for the *Bar* (0.068251) and the least for *Photography Lab* (0.0000042). We list the best five venues in terms of their SS probabilities in table 4. In similar terms, we obtain the SS probabilities for all 247 venues in Tokyo dataset and give the best five venues in terms of SS probabilities in Table 5. Here, we observe that the venue *Train-Station* has the best SS probability (0.34018) among all the venues which makes it the best venue to survive over a long period of time as compared to all other venues in the city. The venue *Afghan Restaurant* (0.0000017) has the least SS probability in Tokyo. This analysis gives the SS probability of the venues present in the two cities which can help in establishing their surviving capacity and popularity over a long period of time. We can observe that *Train-Station* is the single most-popular venue in Tokyo and the popularity distribution in New York is quite evenly distributed among some famous venues.



(a)



(b)

Figure 8. Steady-State probability comparison of venues for New York and Tokyo cities.

We obtained accuracy of 41% to 61% for the venue-prediction, which is comparatively higher than other existing related works, given in section 2. In our work, we also predicted the user presence at a particular venue for the first time, where accuracy is 30%.

Table 4. Top 5 SS probability venues for New York.

Venue Id	Venue	SSP
22	Bar	0.068251
122	Home (private)	0.065552
166	Office	0.054219
224	Subway	0.039614
115	Gym / Fitness Centre	0.039168

Table 5. Top 5 SS probability venues for Tokyo.

Venue Id	Venue	SSP
236	Train Station	0.34018
221	Subway	0.07086
186	Ramen / Noodle House	0.02979
63	Convenience Store	0.028932
128	Japanese Restaurant	0.026902

5 CONCLUSION AND FUTURE DIRECTIONS

LBSNs are used to share locations and events in a participatory fashion. Their popularity is increasing everyday due to the advent of smart-phones. In this paper, mHMM is implemented as a prediction model. Foursquare user check-in datasets for New York and Tokyo cities are used to train the mHMM for qualitative and quantitative evaluation and trend-analysis. The experimental results show that the proposed model performs accurate predictions of user locations. The obtained accuracy of location-prediction of particular venues, in place of categories of LBSNs, using mHMM is very high and outperform the results of various baseline methods. This work will be very useful for *security purposes* in which the exact location of a user, after frequent transitions between different venues, can be traced on a particular day or time. In addition, the restaurant owners can keep track of their *customers' arrivals* on a given day. They get to know their food choice well in advance, that saves time and resources. Also, Steady-State probabilities are obtained, that define the popularity and chance of survival of venues in the future based on the user check-ins. To the best of our knowledge, our work is the first in the domain of location prediction for users on venue basis with such high accuracy results.

In future, we plan to extend the work to capture user behavior in other LBSNs like Facebook, Twitter and Gowalla. Also, we plan to use multi-stage HMM for prediction and compare the results to obtain the best location-prediction. Combination of unstructured information like user-tips along with the temporal activity patterns can result in better predictive performance.

6 ACKNOWLEDGMENT

THIS work was supported by SERB, MHRD, under Grant (EEQ/-2016/000413) for *Secure and Efficient Communications inside Partitioned Social Overlay Networks* project, currently going on at National Institute of Technology Goa, Ponda, Goa, India.

7 DISCLOSURE STATEMENT

NO potential conflict of interest was reported by the authors.

8 REFERENCES

A. Aggarwal, J. Almeida, and P. Kumaraguru. (2013). Detection of spam tipping behaviour on foursquare. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 641-648). ACM.

J. Bao, Y. Zheng, and M.F. Mokbel, (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems* (pp. 199-208). ACM.

L. E. Baum and J. A. Eagon, (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society,* 73(3), 360-363.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics,* 41(1), 164-171.

D. P. Bertsekas and J. N. Tsitsiklis. (2002.). *Introduction to probability*, (2nd edn.), Belmont, MA: Athena Scientific.

J. A. Bilmes. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute,* 4(510), 126.

X. Cao, G. Cong, and C. S. Jensen, (2010). Mining significant semantic locations from GPS data. *Proceedings of the VLDB Endowment,* 3(1-2), 1009-1020.

O. Cappé, E. Moulines and T. Rydén. (2007). Inference in Hidden Markov Models. In *Proceedings of the EUSFLAT Conference* (pp. 14-16). Springer.

C. Cheng, H. Yang, I. King, and M. R. Lyu. (2012). Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks. In *Aaai* (Vol. 12, pp. 17-23).

P. Coppens, C. Veeckman, and L. Claeys. (2015). Privacy in location-based social networks: privacy scripts & user practices. *Journal of Location Based Services,* 9(1), 1-15.

R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo. (2014). Predicting activity attendance in event-based social networks: Content, context and social influence. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 425-434). ACM.

G. Farrelly. (2014). Irreplaceable: the role of place information in a location based service. *Journal of Location Based Services,* 8(2), 123-132.

G. D. Forney Jr. (2005). The viterbi algorithm: A personal history. *arXiv preprint cs*/0504020.

Foursquare stats, 2017, *http://expandedramblings.com/index.php/by-the-numbers-interesting-foursquare-user-stats/.*

Foursquare check-ins pulse, 2017, *https://foursquare.com/infographics/pulse.*

F. Gasparetti. (2017). Personalization and context-awareness in social local search: State-of-the-art and future research challenges. *Pervasive and Mobile Computing,* 38, 446-473.

S. Gambs, M. O. Killijian, and M. N. del Prado Cortez. (2012). Next place prediction using mobility markov chains. In *Proceedings of the*

First Workshop on Measurement, Privacy, and Mobility (p. 3). ACM.

H. Gao, J. Tang, and H. Liu. (2012). Exploring social-historical ties on location-based social networks. In *Icwsm*.

H. Hou, L. Jin, Q. Niu, Y. Sun, and M. Lu. (2011). Driver intention recognition method using continuous hidden markov model. *International Journal of Computational Intelligence Systems,* 4(3), 386-393.

T. Lane. (1999), Hidden Markov Models for Human/Computer Interface Modeling. In *Proceedings of IJCAI-99 Workshop on Learning about Users* (pp. 35-44). Citeseer.

J. Li and L. Li. (2014). A Location Recommender Based on a Hidden Markov Model: Mobile Social Networks. *Journal of Organizational Computing and Electronic Commerce*, 24(2-3), 257-270

I. Litou, I. Boutsis, and V. Kalogeraki. (2017). Efficient techniques for time-constrained information dissemination using location-based social networks. *Information Systems,* 64, 321-349.

Z. Mao, Y. Jiang, G. Min, S. Leng, X. Jin, and K. Yang. (2017). Mobile social networks: Design requirements, architecture, and state-of-the-art technology. *Computer Communications,* 100, 1-19.

W. Mathew, R. Raposo, and B. Martins. (2012). Predicting future locations with hidden Markov models. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 911-918). ACM.

A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. (2011). An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11, 70-573.

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

V. Raghavan, G. Ver Steeg, A. Galstyan, and A. G. Tartakovsky. (2014). Modeling temporal activity patterns in dynamic social networks. *IEEE Transactions on Computational Social Systems,* 1(1), 89-107.

G. Sun, Y. Xie, D. Liao, H. Yu, and V. Chang. (2017). User-defined privacy location-sharing system in mobile online social networks. *Journal of Network and Computer Applications,* 86, 34-45.

C. Sutton and A. McCallum. (2006). An introduction to conditional random fields for relational learning (Vol. 2). *Introduction to statistical relational learning.* MIT Press.

K. Thilakarathna, S. Seneviratne, K. Gupta, M. A. Kaafar, and A. Seneviratne. (2017). A deep dive into location-based communities in social discovery networks. *Computer Communications,* 100, 78-90.

C. R. Vicente, D. Freni, C. Bettini, and C. S. Jensen. (2011). Location-related privacy in geo-social networks. *IEEE Internet Computing,* 15(3), 20-27.

Y. Xiao, X. Lu, and Y. Liu. (2016). A parallel and distributed algorithm for role discovery in large-scale social networks. *Intelligent Automation & Soft Computing,* 22(4), 675-681.

D. Yang, D. Zhang, B. Qu, and P. Cudré-Mauroux. (2016). PrivCheck: privacy-preserving check-in data publishing for personalized location based services. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 545-556). ACM.

D. Yang, D. Zhang, Z. Yu, and Z. Wang. (2013). A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 119-128). ACM.

D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 45(1), 129-142.

J. Ye, Z. Zhu, and H. Cheng. (2013). What's your next move: User activity prediction in location-based social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 171-179).

M. Ye, P. Yin, and W. C. Lee. (2010). Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 458-461). ACM.

M. Ye, P. Yin, W. C. Lee, and D. L. Lee. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 325-334). ACM.

B. Zhao, J. He, Y. Zhang, G. Liu, P. Zhai, N. Huang, and R. Liu. (2016). Dynamic trust evaluation in open networks. *Intelligent Automation & Soft Computing,* 22(4), 631-638.

L. Zhao, Y. Lu, and S. Gupta. (2012). Disclosure intention of location-related information in location-based social network services. *International Journal of Electronic Commerce,* 16(4), 53-90.

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

10 NOTES ON CONTRIBUTORS

Mr. Ahsan Hussain is a Research Scholar in the Department of Computer Science and Engineering, National Institute of Technology Goa, India. His research interests are Data Mining, Computer Networks and Social Media Mining.

Dr. Bettahally N. Keshavamurthy is currently working as an Assistant Professor in the Department of Computer Science and Engineering, National Institute of Technology Goa, India. His research interests include Data Mining, Stream Data Mining, Privacy Preserving Data Mining and Social Media Mining.
Email: bnkeshav.fcse@nitgoa.ac.in

Dr. Ravi Prasad K Jagannath is currently working as Assistant Professor of Mathematics in the Department of Humanities and Sciences, National Institute of Technology Goa, India. His research interests include Nonlinear Optimization, Inverse Problems, and their applications.
Email: k.j.raviprasad@nitgoa.ac.in