Crowd Location Forecasting at Points of Interest

Jorge Alvarez-Lozano

Computer Science Department,
CICESE Research Center
Carr. Ensenada-Tijuana 3918
Ensenada, México
E-mail: jalvarez@cicese.edu.mx

J. Antonio García-Macías*

Computer Science Department,
CICESE Research Center
Carr. Ensenada-Tijuana 3918
Ensenada, México
E-mail: jagm@cicese.mx
*Corresponding author

Edgar Chávez

Institute of Mathematics, UNAM
México
E-mail: elchavez@gmail.com

Abstract: Predicting the location of a mobile user in the near future can be used for a large number of user-centered ubiquitous applications. This can be extended to crowd-centered applications if a large number of users is included. In this paper we present a spatio-temporal prediction approach to forecast user location in a medium-term period. Our approach is based on the hypothesis that users exhibit a different mobility pattern for each day of the week. Once factored out this weekly pattern, user mobility among points of interest is postulated to be markovian.

We trained a hidden Markov model to forecast user mobility and evaluated our approach using a public dataset. The experimental results show that our approach is effective considering a time period of up to seven hours. We obtained an accuracy of up to 81.75 % for a period of 30 minutes, and 66.25 % considering 7 hours.

Keywords: Data mining; Data sharing; Spatio-temporal crowd location forecasting; User location predictability; User mobility similarity.

Reference to this paper should be made as follows: Alvarez-Lozano, J., García-Macías, J. A. and Chávez, E. (2013) ‘Crowd Location Forecasting at Points of Interest’, *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Jorge Alvarez-Lozano is a PhD student in the Computer Science Department at CICESE Research Center, Ensenada, Baja California, México. His research interests include participatory sensing, opportunistic communications, data mining, location-based services, and social networking. Alvarez-Lozano received an MSc in Computer Science from CICESE.

J. Antonio García-Macías is a researcher in the Computer Science Department at CICESE Research Center, Ensenada, Baja California, México. His research interests include ubiquitous computing, human-computer interaction, and computer networks. He received a PhD in Computer Science from Institut National Polytechnique de Grenoble, France.

Edgar Chávez. He received the B.S. degree in Mathematics from the Universidad Michoacana de San Nicolas de Hidalgo, a M.S. degree in Computer Science from the Universidad Nacional Autónoma de México, and a PhD in Computer Science from the Centro de Investigación en Matemáticas, México. He is currently a full professor in Universidad Nacional Autónoma de México. He has chaired major conferences in computer science, has been president of the Mexican Computer Science Society, and have published over 100 technical contributions in different venues. His research interests include storage, retrieval, analysis and mining of data.

1 Introduction

The ability to forecast user location in an accurate way is central to many research areas such as urban planning, healthcare, pervasive and ubiquitous systems, computer networks and recommender systems, to name a few. With the increasing proliferation of mobile devices and the huge variety of sensors incorporated on them, it is possible to register the user location on the move and hence mobile devices become a very rich source of contextual data. We are interested in predicting user location based on her past history, in particular we are interested in knowing if a given user will be at a point of interest at a given time. User mobility patterns have been studied, and researchers have found that people exhibit a high degree of repetition, visiting regular places during their daily activities (Gonzalez et al. (2008)). This regularity of the past movements has been exploited to forecast the next location of the user. This approach has been useful for domains where the time lapse between locations is not relevant; a more complex scenario require also to know *when* the user will be at the given location. The former problem has been analyzed in Scellato et al. (2011) and Sadilek & Krumm (2012), they are able to predict the user location in space and time (a.k.a. *spatio-temporal prediction*) for short and long-term time lapses respectively. We are interested in a more precise forecasting of the user location, let's see the scenario. Assume current time is 10:00 AM (T), what we want to know is: Where a user will be in the next 3 to 5 hours ($[T, T + \Delta(T)]$ for some $\Delta(T)$). We have not found in the literature a solution for this medium-term spatio-temporal prediction.

After mining individual mobility patterns, we are able to join the individual predictions, and to serve unspecific applications for the crowd. The general public (including banks, retail stores or traffic authorities) could query a location forecasting service to know the capacity of a specific Point of Interest (POI). In one application scenario, a user may query the capacity of a given POI (think for example in a fast food location) and decide between many different alternatives based on that measure.

A more specialized service can warn the user about a potential overcrowded POI she is about to visit in the next few hours. This is a potential paradox, since all users using the service may back-off and avoid the POI, which will not be overcrowded at the end. This may have the same effect of trying to predict the behavior of individual stock options. Any successful forecasting algorithm can change the behavior of a system. For the time being we will focus on forecasting the user location without worrying about paradoxical behavior.

Users should be assured about the potential usage of their data. Disclosing their current location implies they are disclosing travel patterns and the individual forecast of their location. The service should provide secure ways to disclose the location of anonymous users, since the

objective is to infer the location of a crowd in a given POI.

In this work, we present a medium-term spatio-temporal prediction model based on the repetitive patterns of people visiting some specific places that are important for them. We found that the user mobility among the POIs can be modeled as a Markovian chain. Then, considering the Markovian property among POIs, and the relation of staying time with specific time periods, we modeled the mobility as a hidden Markov Model (HMM).

The proposed approach to forecast the crowd location consists of two steps. First, for each user we identify the current significant places or points of interest. Then, using the records of the visits to the significant places, we define the user mobility as an HMM, with which we are able to predict where a user will be in a given time period. The contribution of this paper can be summarized as follows:

- We identified the minimum time period covering the current mobility pattern of a user. We compared the daily user mobility using the cosine similarity and as a result, we identified current points of interest. This enabled an accurate prediction model.
- We present a method to forecast the user location using a hidden Markov model. Our approach postulate, and experimentally prove, a markovian property found in the user mobility among POIs.
- We used the public GeoLife GPS trajectory dataset (Zheng et al. (2008, 2009, 2010)) to evaluate our approach.
- We compared our approach with a method based on NextPlace (Scellato et al. (2011)). With our method, we report an overall prediction accuracy of up to 85 % for short prediction periods, with an accuracy of up to 70 % when the prediction period is 7 hours, both higher than the baseline of NextPlace.

The rest of this paper is organized as follows: we start discussing some of the work related to spatial and spatio-temporal forecasting, and discuss the characteristics of the mobility defining the spatio-temporal prediction model. Later we present a methodology to define the spatio-temporal prediction, introducing some application scenarios, and proceed with an experimental evaluation using a public data set. Finally, we discuss the scope and limitation of our approach and give some conclusions.

2 Related work

Most of the current methods for mining and forecasting user location are oriented to a single user instead of a crowd. In applications where a crowd is going to be

tracked at once the method should scale. We believe that using an automata for each user to predict her position is an scalable way to implement crowd forecasting because the automata related to an user can be queried in constant time. This is the motivation of our work and our solution approach.

Mobility prediction, mobility mining or mobility forecasting have many different sources depending on the application. Some authors try to characterize activities and traveling paths to obtain behavior patterns and to predict user mobility (Ashbrook & Starner (2003), Chon et al. (2012), Eagle & Pentland (2006), Kim et al. (2006), Sadilek & Krumm (2012), Ye et al. (2009), Yuan et al. (2011), Zheng et al. (2008), Zheng & Xie (2010)). According to our approach, we can distinguish the related work as *spatial prediction and spatio-temporal prediction*.

The goal of the spatial prediction is to identify the next place or places where the user will be. Some existing approaches are able to forecast just the next location of a user, but not the time when the user will be at the place (Ashbrook & Starner (2003), Kim et al. (2006), Krumm & Horvitz (2006), Liao et al. (2006), Nicholson & Noble (2008), Song et al. (2006)). Krumm & Horvitz (2006) present a method to forecast the driver's destination using GPS trajectories as the journey progresses and the driver moves from a geographical region to another. A similar approach is presented in Yavas et al. (2004) where the objective is to predict in which cell of a cellular network the user will be, being the restriction the topology, there are only a limited number of cells the user can reach from the current location. Another work is BreadCrumbs (Nicholson & Noble (2008)), the authors use discrete data to estimate the next user location providing network connectivity. In WhereNext (Monreale et al. (2009)), the authors use previous trajectory patterns (sequences of regions), which are modeled as a tree. Later, considering a new trajectory the next location is predicted finding the most likely sequence path in the tree. Using the above approaches, predictions cannot be extended into the future because they are just focused on the next place to be visited. These previous works are based on markovian or bayesian models. A more recent work uses HMMs, and the same dataset we use to forecast the user location. However, they define on a different way the HMM; they are interested in the mobility among geographical regions instead of points of interest (Mathew et al. (2012)). An evaluation of several techniques to forecast the next user location can be found in Song et al. (2003).

Regarding to spatio-temporal prediction, in Vu et al. (2011) traces of Wi-Fi and Bluetooth (the user log to a place) are used to find regular patterns in order to predict where a user will be at a given time, how long the user will stay at that place, and also infer who the user will meet there. To do the spatio-temporal prediction, traces are split according to different time periods. Then, for a given period the most likely location is returned. Meanwhile, Krumm & Brush (2011) present a method

to predict when a user will be at home or away, based on previous observations of presence there. One particular approach is NextPlace (Scellato et al. (2011)), which is general enough to be significant to compare with our approach. The authors get as reference the sequence of the current places visited by a user, and using the historical records search for a similar pattern in the past to predict the next place the user will visit and the time she will stay there. In Scellato et al. (2011), the best accuracy (about 90 %) is obtained when the prediction period ($\Delta(T)$) is only about 5 minutes, if the prediction period increases to 60 minutes the accuracy drops to 70 %. Although the above works have been useful in some domains, they do not offer the functionality required to do the medium-term spatio-temporal prediction. Also, they do not consider just the current mobility pattern, neither they distinguish a particular pattern for each day of the week. Finally, they do not consider the transition among significant places.

3 User mobility

Although user mobility seems to be dynamic, most people follow certain mobility patterns (Eagle & Pentland (2006), Farrahi & Gatica-Perez (2011), Gonzalez et al. (2008)); it is rare to have a completely erratic behavior over time. Mobility is fixed by our activities and habits, like working, school attendance, recreational endeavors, and other activities that vary over time within certain behavior boundaries. We can distinguish between weekday, weekend, monthly or annually patterns. This is analyzed in Eagle & Pentland (2006), Motahari et al. (2012), Yavas et al. (2004). Once recognized the user mobility patterns, we are able to predict her spatio-temporal location. In this work we fixed our attention in three major categories of user mobility and with their interrelation we can infer the user mobility in general, Figure 2 illustrates.

3.1 Temporal patterns

It is reasonable to assume certain periodicity in location/time patterns. Usually weekdays are similar, people tend to organize their life according to work or school hours, and our hypothesis is that activities in the same weekday will have a repetitive pattern. A corresponding periodicity is observed during weekends. Mobility exhibits a different pattern for each day of the week (Chon et al. (2012), Farrahi & Gatica-Perez (2011), Hsu et al. (2007), Motahari et al. (2012)). Hence, the places visited in a given day are postulated to be the same for the subsequent days.

Two additional observations are that the current location in a given day and hour conditions the next place to be visited. For example if one user is at home at 7:00 AM on a Monday, the next place he will be at is most likely the coffee shop or the office, but not the restaurant or a movie theater. We postulate the

user mobility is a markovian stochastic process and can be described with a Markov Chain. The Markovian property (Markov (1961)) would state that current place is only a function of the previous place. Our main claim is that once the data is grouped by day and time, the sequence of places visited form a Markovian chain.

3.2 Markovian chain among POIs

Considering our above claim, it is necessary to clarify that the markovian property is just valid when the user moves among certain discrete places. These discrete locations are distinguished because the user spends some time and visit them frequently; they are known as significant places or *points of interest* (POI). For instance, considering the user mobility on a given Monday, after he is at home, he goes to the coffee shop, then to a super market, then to his work. After some hours, he goes to the restaurant, returns to his work, then he goes to the gym, and finally he goes back home. Next Monday, his mobility does not differ too much, after leaving home, he goes to the coffee shop, then to the laundry, and work. Around 2:00 PM, he goes to the restaurant, returns to his work, then he goes to the gym, and finally he goes back to home. Considering this scenario, it is possible to identify places that are POIs (Figure 1 black shape), and those which visit occasionally (Figure 1 dotted line). Also, we can observe that the mobility among POIs does not change (home, coffee shop, work, restaurant, work, gym, home); just the places visited in the transition between POIs change. We refer the reader to the Figure 1.

3.3 Modeling user mobility as a hidden Markov model

Using the above hypothesis, it is possible to model the user mobility as an HMM (Rabiner (1989)) (See Figure 5). A hidden Markov model is a finite state machine consisting of a set of hidden states (Q), a set of observations (O), transition probabilities (A), emission probabilities (B), and initial probabilities for each state (π). Hidden states are not directly visible, while observations (dependents on the state) are visible. Each state has a probability distribution over the set of observations ($\lambda = (A, B, \pi)$). In our case, the hidden states corresponds to POIs, which have a probability distribution over times of day. Once the HMM has been defined, three inference problems can be addressed, (i) finding the probability of an observed sequence (evaluation); (ii) finding the sequence of hidden states that most probably generated an observed sequence (decoding); (iii) generating an HMM given a sequence of observations (learning). For the purpose of this work, we use the *decoding approach*; given a time period (sequence of observations), we want to know the most likely sequence of locations where the user will be (hidden states). More detail can be found in section 4.5.

4 Methodology

To address the challenge of forecasting user mobility in space and time, we first convert the GPS readings into indoor and outdoor POIs. Then, using the records of the visits to the identified POIs, an HMM is defined for each day and for each user. Once, the HMM is defined, we make some predictions.

4.1 Identifying points of interest

Discovering significant places for an user has been an important research topic in several applications such as ambient assisted living, ubiquitous computing, and others. Mobile traces produced by the mobile devices, provide a great amount of location data useful to discover where the user spends her time. Thereby, with this amount of location data there is a need for algorithms that deal with the challenge of turning data into significant places (Ashbrook & Starner (2003), Kang et al. (2005), Kim et al. (2006), Marmasse & Schmandt (2000), Palma et al. (2008), Ram et al. (2010), Zheng et al. (2009), Zhou et al. (2007)). At the moment, some works have been focused on discovering these places using different approaches, and for different purposes. These can be classified as follows:

- Residence time-based algorithms. This approach is based on the assumption that the importance of a place is directly proportional to the permanence at it (Kang et al. (2005), Kim et al. (2006), Scellato et al. (2011), Ye et al. (2009)).
- Density-based clustering algorithms. These are based on the density of GPS points (Ram et al. (2010), Zhou et al. (2007)).
- Lost signal-based algorithms. This approach has been used to discover indoor POIs, using the GPS signal disappearance and reappearance (Ashbrook & Starner (2003), Marmasse & Schmandt (2000)).

The intuition behind the first two approaches offers a good functionality for some specific applications, nonetheless, for our approach they have a limitation. It is not possible to take into account just one of those, because maybe a given user was at a place for a long time, but that happened some months ago; or, considering the density-based algorithms, a place can be considered significant after taking into account the GPS points of several visits. Thereby, it is necessary to consider time and location data, as well the frequency of visits and the time period that covers the current mobility pattern. This way, we are enabled to identify the places that are currently significant, and therefore, define in a better way the forecasting model.

In order to identify indoor and outdoor POIs, we have applied the below criterions to the algorithms of previous works (Ashbrook & Starner (2003), Kang et al. (2005)).

- Residence time.

Figure 1 Some significant places are found in the daily user activities.

Figure 2 The next user location depends only on the current location, once the sequences have been factorized by week day and time of the day.

- Cluster radius.
- Frequency of visits.
- Time period (Windows size).

The Ashbrook & Starner (2003)'s algorithm focuses on discovering indoor significant places. To do that, they consider the GPS loss signal within a fixed radius, and a time threshold for the disappearing period. Kang et al. (2005) propose a time-based clustering algorithm to discover outdoor significant places. They compare each incoming GPS reading with previous readings in the current cluster; if the stream of readings moves away from the current cluster, then they form a new one. Kang et al. (2005) consider two thresholds, d and t for distance and staying time respectively. If the GPS readings are close together (within some distance d of each other), and the user spends at least t minutes at that region, a cluster is formed. When the user moves away from the current cluster, a new cluster is formed; the cluster is discarded if the user stays less than t time.

The above variables allow us to identify the places where the user spends some time and visits frequently. Also, considering the cluster radius we can identify POIs in different levels. For instance, home can be considered as a granular POI, and the shopping mall as a zone POI. Thus, it is possible to forecast the user location in two levels. The last variable is very important to define an accurate forecasting model, and it is necessary to answer the question: *What time period defines in a better way the current mobility pattern?* As we mentioned before, the user mobility varies over time; it is important to identify the current mobility pattern and the time windows size in which this pattern has been in place. Thereby, we can identify just the current POIs. That

is, if the current day is a Monday i , the challenge is to determine the quantity of previous Mondays that are similar in terms of mobility.

To identify the time period that covers the current mobility pattern, we used the cosine similarity; comparing the Day_i vector with the Day_{i-1} vector. The window size increases if the similarity is above some threshold Θ ; otherwise, skip the records of the Day_{i-1} , and compare Day_i vector and Day_{i-2} vector. If the similarity of m consecutive days is below some threshold Θ , the window size ends, and just includes the records of the days with a similar mobility.

We used multiple values for residence time and cluster radius. According to Ashbrook & Starner (2003), the minimum residence time (t) would be 10 minutes; we also consider 30, and 60 minutes. For the cluster radius, we defined 5 values (5, 25, 50, 75, 100 meters); for the frequency, a place must have at least n visits, where $n = \text{window size} / 2$. Using the set of values for the above variables, we combined them in order to know the best values for each variable. In section 6, we present the results.

It is important to notice that even if we identify indoor and outdoor POIs, when we make the predictions we do not make differences between them.

4.2 Converting user mobility into a vector

In order to compare the user mobility by day, each day has been converted into a vector. For each day, we divide it into 48 periods of 30 minutes; each slot contains an index (starting from 1) that corresponds to a POI where the user has been in that period, as shown in Figure 3. The index 0 defines an *unknown place*; a user is at a non-POI or she is moving. We defined the size of the slot to 30 minutes in order to know with this level of granularity

where the user has been; we also considered slots of 15 and 60 minutes; however, there is no significant difference considering a slot size of 15 minutes. In contrast, using a slot of 60 minutes we obtained a lower similarity; it was more likely to find some POIs in a same slot.

4.3 Updating POIs

As mentioned before, the user activities vary over time, and therefore his mobility. Consider for example that a given user is starting swimming classes on Monday, from 7:30 AM to 8:30 AM at a public pool (Top of Figure 4, block D). For the subsequent Mondays, the user takes his classes at the same time; however, the public pool is not a POI yet because it does not have the number of visits required (the relative frequency is small). Some weeks after, the public pool has the required visits to be a POI; the HMM is updated (Bottom of Figure 4). The HMM also is updated when a place ceases to be a POI. For instance, when the mentioned user no longer takes his swimming classes. As Ashbrook & Starner (2003) mentioned, a limitation of using the Markov approach is that a behavior change may take a long time to be reflected. To address this challenge, we propose to update day after day the POIs and the information related to them (Figure 4). A feasible option would be that the user explicitly specifies his behavior change, or that the user selects from a set of recognized patterns, the appropriate one according to the season.

4.4 Defining user prediction model

The HMM is defined as follows:

- Hidden states. These are defined by the POIs. Also, another hidden state was added to define when the user is at a location that is not a POI.
- Observations. These are defined by the average of the arrival and leaving time to the POIs. According to Do & Gatica-Perez (2012), the arrival and leaving times to some places do not change much; for instance, considering the work or scholar activities, there are times defined for arrival and leaving. This way, we can define in an accurate way the time when the user will be at a POI; otherwise, considering the leaving time, we can define that he is at another POI, or at a non POI.
- Vector π . It defines the probability that the user starts his day at a given POI.
- Transition matrix. It defines the probability that the user moves from a POI to another, or from a POI to the state that corresponds to a non-POI.
- Confusion matrix. It defines the probability that the user is at a given POI (or at a non-POI), at a given time.

Once the HMM is defined, we are able to forecast the user location in a given time period (See Figure 5). For instance, if the current time is 11:00 AM, and we want to know where the user will be in the period 11:00 - 18:00, the vector π , transition matrix (A), and confusion matrix (B), are used to identify the combinations of hidden states (and their corresponding probabilities) that satisfy such time period, so later the sequence of hidden states with the highest probability can be selected (*decoding problem*). In order to identify the sequence of hidden states with the highest probability, we used the Viterbi algorithm (Viterbi (2006)).

4.5 Viterbi's algorithm

As mentioned on section 3.3, the aim of the decoding problem is to discover the hidden state sequence that was most likely to have produced a given observation sequence. One solution is to use Viterbi's algorithm to find the single best state sequence for an observation sequence. Viterbi's algorithm is a trellis algorithm that is very similar to the forward algorithm used in the *evaluation problem*, except that the transition probabilities are maximized at each step, instead of added (See Figure 6). Viterbi's algorithm is as follows, first it defines:

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(q_1 q_2 \dots q_t = s_i, o_1, o_2 \dots o_t | \lambda) \quad (1)$$

as the probability of the most probable state path for the partial observation sequence. Then:

Initialization:

$$\delta_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N, \psi_{1(i)} = 0 \quad (2)$$

Recursion:

$$\delta_t s(i) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), \quad 2 \leq t \leq T, 1 \leq j \leq N \quad (3)$$

$$\psi_t(j) = \operatorname{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}], \quad 2 \leq t \leq T, 1 \leq j \leq N \quad (4)$$

Termination:

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (5)$$

$$q_T^* = \operatorname{argmax}_{1 \leq i \leq N} [\delta_T(i)] \quad (6)$$

Optimal state sequence backtracking:

$$q_t^* = \psi_{t+1} q_{t+1}^*, t = T-1, T-2, \dots, 1 \quad (7)$$

5 Evaluation

To test the efficacy of our spatio-temporal prediction, we compared our approach against a modified version of NextPlace (hereby defined as NP*). To the best of our knowledge, NextPlace's results have been the best

Time \ Week	...	07:30	08:00	08:30	09:00	...	23:30
1	...	1	0	0	2		1
2	...	0	0	0	2		1
3	...	0	0	2	2		1

Figure 3 User mobility as a vector. **Figure 4** The update process of POIs is made for each day of the week (e.g. Monday) considering the user mobility week over week **Figure 5** An HMM representing some POIs and their relationship with different hours of the day.

in the literature. NextPlace includes two methods: one to predict the user location at a given time, and another to predict the residence time at that place. The NP* method just includes the method to forecast the user location at a given time; POI selection remains the same. Our method estimates the full sequence of POIs where the user will be in a given time period, while NP* estimates the place where the user will be at a specific time. For this reason, and to do a fair comparison of the performance of both methods, for each observation in the time period we apply the NP* method using the best

value for m ($m = 3$, more detail about the variable m can be found in Scellato et al. (2011)) to predict where the user will be. That is, if we want to know the sequence of POIs in the $[11:00, 15:00]$ period, and considering that for this period there are 3 observations defined (e.g. 11:30, 13:00, 14:30), we apply the NP* method for each one of these observations. Then, we average the results in order to compare them with our method.

using the cosine similarity. If the similarity is above some threshold Θ ($\Theta = 0.50$), the current HMM is updated with the records of the predicted day. Otherwise, the current HMM is used to predict the user location for the test week 2 (i.e. 9th Monday). This process is applied for the subsequent weeks 3 and 4 (i.e. 10th and 11th Monday)

For each spatio-temporal prediction model, we have made five predictions considering different values for ΔT (30 minutes, 1, 3, 5, and 7 hours). A total of 140 predictions for each user, 35 for each test week. Each prediction is uniformly distributed over the [00:00,17:00] interval. We selected this interval because if $T = 17:00$, it is possible to know in which POIs a user will be using the largest value for ΔT (7 hours). In total, 8820 predictions were made using 1st order HMM.

5.4 Effectiveness of the prediction model

To determine the effectiveness of the prediction, if we estimate where a user will be in the interval $[T, T + \Delta T]$, the prediction is correct if the user is at place q_i in the interval $[T_{pred} - \theta, T_{pred} + \theta]$, θ represents an error margin. That is, a prediction is correct when the user is at the POI defined by q_i , at the time indicated by the observation o_i with certain error margin. It would also be correct if the prediction indicates that the user is not at a POI (in the case of the state corresponding to an unknown place). We have defined $\theta = 15$ minutes.

$$T_{pred} = T + o_i \quad 1 \leq i \leq \text{No. of obs. in the interval} \quad (8)$$

6 Results

6.1 POIs

To define an accurate prediction model, we identified the time period that comprises the current mobility pattern. Thus, for each user, and for each day of the week, we identify this period. On average the time period (windows size) is about 7.3 weeks, the shorter time period is 4 weeks, and the longer is 11 weeks; the time period is used for training each HMM.

We want to find the values of minimum residence time and cluster radius maximizing POI identification. Figure 7 show the amount of time the users spend at POIs considering 3 values for residence time (10, 30, and 60 minutes), and five values for the cluster radius (5, 25, 50, 75, 100 meters). As it can be seen in Figure 7, the minimum residence time does not increase significantly POI identification ($\approx 1\%$) even considering a cluster radius of 50 meters. In Figure 8 (Left) we compare the percentage that the users spend at POIs considering the cluster radius of 50 meters and the discovered minimum residence times. The time spent at POIs is maximized when a residence time of 10 minutes is considered ($\approx 55.14\%$); in terms of hours this percentage indicates that users spent 13.23 hours at POIs (Figure 8 right).

Table 1 Number of POIs identified per user

	Min	Max	Mean
Pois	1	7	3.69
Pois on weekdays	2	7	3.91
Pois on weekends	1	6	3.16

Maximizing the percentage, we ensure the identification of the most important and significant places for the user, enabling us to define a better prediction model. Even though the average percentage that the user spent in POIs is about 55 % (just the 43% of the trajectories have a duration longer than an hour), we obtained good results. It would be desirable to continuously collect the user location as much as possible to know where the user spends his time, and to define the prediction model accurately.

On average each user has 3.69 POIs; the minimum number of POIs found per day was 1, and the maximum was 7. The average number of POIs corresponds to what was commented in Chon et al. (2012), a high degree of regularity is found in a few places. Grouping POIs by weekdays and weekends, on average people have more POIs on weekdays (3.91); in contrast, people have on average 3.16 POIs on weekends (Table 1). Finally, on average the highest number of POIs was found on Tuesdays (4.11), and the lower number was found on Sundays (3.00).

6.2 Predictability

Respecting to the mobility predictability, for each day and for each user, we applied the predictability test. The significance level α was set to 0.05, as defined in Zhang et al. (2010). According to the results, on average for each user, 60.31 % of his prediction models (this is equivalent to ≈ 4.22 days) are predictable; the predictions models corresponding to weekends are more predictable than weekdays, 66.00 % in contrast to 60.00 %. The day with the lowest percentage of predictability was Tuesday, just 33.33 % of its prediction models were predictable; in contrast, 77.77 % of the prediction models for Friday were predictable. Finally, the lowest percentage for a user was 14.28 % (≈ 1 day), and the highest value was 100 % (7 days).

6.3 Prediction

We present the prediction results in two parts. First, in Figure 9 we present the accuracy obtained for the different test weeks. Figure 9 a) presents the results obtained when we predict the user location for the test week 1; for the second test week the 38 % of the HMM models were updated, the results are presented in b); in c) we present the results obtained for the third test week, where the 63 % of HMM models were updated; finally, the results for the 4th week are presented in d), the 24 % of HMM were updated. We obtained results of up to

Figure 7 Comparing different cluster sizes and residence time. Percentage of the day spent at POIS considering different cluster sizes and minimum residence times.

Figure 8 Comparing residence times using cluster radius of 50 meters. Comparing the percentage of the day (left) and hours (right) spent at POIs when we consider a cluster radius of 50 meters and different minimum residence times.

85 %, 82 %, 78 %, 75 %, 70 %, for the time periods of 30, 60, 180, 300 and 420 minutes, respectively.

In Figure 10, we show the average results of the four test weeks. When considering a period of 30 minutes, we obtain an accuracy of 81.75 % against 74 % for NP*; for a period of 60 minutes, we obtain 79 % against 66.75 % for NP*. When the value of ΔT increases, the accuracy for NP* decreases more than our model. For periods longer than 60 minutes, the accuracy for NP* lies within the range of 45 % - 65 %. In contrast, with our proposal, we obtain an accuracy of 75 % for a period of 3 hours; a period of 5 hours yields an accuracy of 72 %, and finally a 7 hours period yields an accuracy of 66.25 %. However, for periods beyond 7 hours, the accuracy obtained drops down to 40 %. The previous evaluation shows that our proposal can yield good accuracy to predict the user mobility in a medium-term up to 7 hours. Furthermore, we show that considering the three characteristics of mobility (week day, time of the day, and current location) and their relationship, allow us to have a more accurate prediction model, and better results as consequence. We show that our method is better than the NP* method.

The fact that our approach get better results than NP*, can be explained by some factors: (i) POIs selection on NP* is based just on the residence time. This way important places with less residence time and high number of visits are discarded, and as a result a non accurate model is defined; (ii) also, this method does not take into account the temporal mobility pattern. For the POI selection and training process, NP* does not consider the current mobility behavior; (iii) this method does not define a model for each day of the

week; just one spatio-temporal prediction model is used for all predictions. If we want to get good results, a prediction model for each day must be defined, because as mentioned in Chon et al. (2012), Hsu et al. (2007), Motahari et al. (2012) user mobility is different for each day of the week; (iv) finally, this method does not consider the transition among POIs, it is just focused on the arrival and residence time at POIs.

7 Application scenarios

Once we have knowledge about the user mobility, it is feasible to join the predictions of the population to create some interesting applications for the crowd. In this work we have focused on scenarios where the participation of the population is required, and thus bringing them benefits. We present some possible scenarios where the medium-term prediction is required, and we can address them with our approach.

7.1 Avoid congestion locations

Consider that on a given Friday, John and his wife go to dinner to a famous restaurant. When they arrive to the place, there are no locations available, and several people are waiting for a place. After a while, they are annoyed and decided to go to another place. This situation can be avoided, if crowd location forecasting is taking into account. This way, each user can share her location forecasting every day, in order to estimate the number of persons in a given place at a given time. As a result,

Figure 9 Prediction accuracy using different values for ΔT

Figure 10 Prediction accuracy using different values for ΔT

John and his wife, could have queried the desired place in order to know how many people there will be, and considering the capacity of this place, decide if there will be an available place; otherwise, they could look for another place (See Figure 11).

7.2 Reserve resources

At the moment, some works have used the prediction of the next location(s) to reserve resources (Nicholson & Noble (2008)); for instance, bandwidth at a given hotspot. Although the information about the next

location is important to do an accurate reserve, a better reserve can be done when the arrival time to the next location(s) is considered. This way, before the user arrive to the place, the resources can be assigned or used by another user.

8 Conclusions and future work

In this work we take advantage of the Markovian property shown in the mobility behavior of most people in their everyday activities. However, we acknowledge

Figure 11 A possible mobile application to know how many people there will be in a given time at a given location.

that there is a minority of individuals that do not follow predictable patterns (e.g. politicians, CEOs, etc.) and have constantly changing agendas; these individuals would require another type of mechanism to forecast their mobility. Even in the case where mobility displays the Markovian property, there may be exceptional situations where individuals deviate from their normal behavior, for instance on mother's day or some special anniversary. In such special occasions, it might be necessary to consider applying a different prediction model, like the use of eigen-behaviors (Eagle & Pentland (2009)). These special occasions could be inferred by obtaining personal context from different sources, such as the phone's calendar or social networks events (e.g. Facebook, Google Plus).

There are some aspects that we will be addressing in order to improve this work: (i) due to the metrics used on the POI selection process, we identified just a few number of POIs per user, it would be interesting to consider more places, and as a result the population would benefit; (ii) in this work we have used continuous data, as mentioned in a previous section it would be desirable to combine continuous and discrete data to define the prediction model, and then forecasting the user location at granular and zone level. For instance, using GPS and Wi-Fi data, and also social networks presence status (Facebook, Twitter, Foursquare, or Google Plus check-ins).

8.1 Limitations

With our current proposal, it is necessary to join the individual predictions in a centralized entity, to later query it to know how many people will be at a given location at a given time. For this reason, we required to consider security actions to preserve the user's privacy

and anonymity. An interesting and important aspect to define in an accurate way the mobility model, is to identify the time period that covers the current mobility pattern. Also, identify as soon as possible behavior changes to update the respective model. A natural limitation found in this and another works is the quantity of data available for analysis (Ashbrook & Starner (2003)).

References

- Ashbrook, D. & Starner, T. (2003), 'Using gps to learn significant locations and predict movement across multiple users', *Personal Ubiquitous Computing* **7**(5), 275–286.
- Chon, Y., Shin, H., Talipov, E. & Cha, H. (2012), Evaluating mobility models for temporal prediction with high-granularity mobility data, in 'Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications', Percom, IEEE, pp. 206–212.
- Do, T. M. T. & Gatica-Perez, D. (2012), Contextual conditional models for smartphone-based human mobility prediction, in 'Proceedings of the 2012 ACM Conference on Ubiquitous Computing', UbiComp '12, ACM, pp. 163–172.
- Eagle, N. & Pentland, A. (2006), 'Reality mining: sensing complex social systems', *Personal Ubiquitous Computing* **10**(4), 255–268.
- Eagle, N. & Pentland, A. (2009), 'Eigenbehaviors: identifying structure in routine', *Behavioral Ecology and Sociobiol.* **63**(7), 1057–1066.

- Farrahi, K. & Gatica-Perez, D. (2011), 'Discovering routines from large-scale human locations using probabilistic topic models', *ACM Transactions on Intell. Systems and Technology* **2**(1), 3:1–3:27.
- Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. (2008), 'Understanding individual human mobility patterns', *Nature* **453**(7196), 779–782.
- Hsu, W., Spyropoulos, T., Psounis, K. & Helmy, A. (2007), Modeling time-variant user mobility in wireless mobile networks, in 'Proceedings of the 26th IEEE International Conference on Computer Communications', INFOCOM, IEEE, Anchorage, Alaska, USA, pp. 758–766.
- Kang, J. H., Welbourne, W., Stewart, B. & Borriello, G. (2005), 'Extracting places from traces of locations', *SIGMOBILE Mobile Computing and Communications Review* **9**(3), 58–68.
- Kim, M., Kotz, D. & Kim, S. (2006), Extracting a mobility model from real user traces, in 'Proceedings of the 25th IEEE International Conference on Computer Communications', INFOCOM '06, pp. 1 – 13.
- Krumm, J. & Brush, A. J. B. (2011), Learning time-based presence probabilities, in 'Proceedings of the 9th International Conference on Pervasive Computing', Pervasive'11, Springer-Verlag, pp. 79–96.
- Krumm, J. & Horvitz, E. (2006), Predestination: inferring destinations from partial trajectories, in 'Proceedings of the 8th International Conference on Ubiquitous Computing', UbiComp'06, Springer-Verlag, pp. 243–260.
- Liao, L., Patterson, D. J., Fox, D. & Kautz, H. (2006), 'Building personal maps from gps data', *Annals of the New York Academy of Sciences* **1093**, 249–265.
- Markov, A. A. (1961), *Theory of Algorithms*, Israel Program for Scientific Translations, Bloomington, IN, USA.
- Marmasse, N. & Schmandt, C. (2000), Location-aware information delivery with commotion, in 'Proceedings of the 2nd International Symposium on Handheld and Ubiquitous Computing', HUC '00, Springer-Verlag, pp. 157–171.
- Mathew, W., Raposo, R. & Martins, B. (2012), Predicting future locations with hidden markov models, in 'Proceedings of the 2012 ACM Conference on Ubiquitous Computing', UbiComp '12, ACM, pp. 911–918.
- Monreale, A., Pinelli, F., Trasarti, R. & Giannotti, F. (2009), Wherenext: a location predictor on trajectory pattern mining, in 'Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '09, ACM, pp. 637–646.
- Motahari, S., Zang, H. & Reuther, P. (2012), The impact of temporal factors on mobility patterns, in 'Proceedings of the 2012 45th Hawaii International Conference on System Sciences', HICSS '12, IEEE Computer Society, pp. 5659–5668.
- Nicholson, A. J. & Noble, B. D. (2008), Breadcrumbs: forecasting mobile connectivity, in 'Proceedings of the 14th ACM International Conference on Mobile Computing and Networking', MobiCom '08, ACM, pp. 46–57.
- Palma, A. T., Bogorny, V., Kuijpers, B. & Alvares, L. O. (2008), A clustering-based approach for discovering interesting places in trajectories, in 'Proceedings of the 2008 ACM Symposium on Applied Computing', SAC '08, ACM, pp. 863–868.
- Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proc. of the IEEE* **77**(2), 257–286.
- Ram, A., Jalal, S., Jalal, A. S. & Kumar, M. (2010), 'A density based algorithm for discovering density varied clusters in large spatial databases', *Int. J. of Computer Applications* **3**(6), 1–4. Published By Foundation of Computer Science.
- Sadilek, A. & Krumm, J. (2012), Far out: Predicting long-term human mobility, in 'Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence', AAAI, AAAI Press.
- Scellato, S., Musolesi, M., Mascolo, C., Latora, V. & Campbell, A. T. (2011), NextPlace: a spatio-temporal prediction framework for pervasive systems, in 'Proceedings of the 9th International Conference on Pervasive Computing', Pervasive'11, Springer-Verlag, pp. 152–169.
- Song, L., Deshpande, U., Kozat, U. C., Kotz, D. & Jain, R. (2006), Predictability of wlan mobility and its effects on bandwidth provisioning, in 'Proceedings of the 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies', INFOCOM '06, IEEE.
- Song, L., Kotz, D., Jain, R. & He, X. (2003), 'Evaluating location predictors with extensive Wi-Fi mobility data', *SIGMOBILE Mob. Computing and Communications Review* **7**(4), 64–65.
- Viterbi, A. J. (2006), 'A personal history of the Viterbi algorithm', *IEEE Signal Processing Magazine* **23**(4), 120–142.
- Vu, L., Do, Q. & Nahrstedt, K. (2011), Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace, in 'Proceedings of the 2011 IEEE International Conference on Pervasive Computing and

- Communications', PERCOM '11, IEEE Computer Society, pp. 54–62.
- Yavas, G., Katsaros, D., Ulusoy, O. & Manolopoulos, Y. (2004), 'A data mining approach for location prediction in mobile environments', *Data & Knowl. Engineering* **54**(2005), 121–146.
- Ye, Y., Zheng, Y., Chen, Y., Feng, J. & Xie, X. (2009), Mining individual life pattern based on location history, in 'Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware', IEEE Computer Society, Washington, DC, USA, pp. 1–10.
- Yuan, J., Zheng, Y., Zhang, L., Xie, X. & Sun, G. (2011), Where to find my next passenger, in 'Proceedings of the 13th International Conference on Ubiquitous Computing', UbiComp '11, ACM, pp. 109–118.
- Zhang, Y. F., Zhang, Q. F. & Yu, R. H. (2010), Markov property of markov chains and its test, in 'Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)', IEEE, pp. 1864–1867.
- Zheng, Y., Li, Q., Chen, Y., Xie, X. & Ma, W.-Y. (2008), Understanding mobility based on gps data, in 'Proceedings of the 10th International Conference on Ubiquitous Computing', UbiComp '08, ACM, pp. 312–321.
- Zheng, Y. & Xie, X. (2010), Learning location correlation from gps trajectories, in 'Proceedings of the 2010 Eleventh International Conference on Mobile Data Management', IEEE Computer Society, Washington, DC, USA, pp. 27–32.
- Zheng, Y., Xie, X. & Ma, W.-Y. (2010), 'Geolife: A collaborative social networking service among user, location and trajectory', *IEEE Data Engineering Bulletin* **33**(2), 32–39.
- Zheng, Y., Zhang, L., Xie, X. & Ma, W.-Y. (2009), Mining interesting locations and travel sequences from gps trajectories, in 'Proceedings of the 18th International Conference on World Wide Web', WWW '09, ACM, pp. 791–800.
- Zhou, C., Frankowski, D., Ludford, P., Shekhar, S. & Terveen, L. (2007), 'Discovering personally meaningful places: An interactive clustering approach', *ACM Transactions on Information Systems* **25**(3), 56–68.