Predicting Future Locations with Hidden Markov Models

Wesley Mathew
wesley.mathew@ist.utl.pt

Ruben Raposo
ruben.raposo@ist.utl.pt

Bruno Martins
bruno.g.martins@ist.utl.pt

INESC-ID
Instituto Superior Técnico
Av. Professor Cavaco Silva
2744-016 Porto Salvo,
Portugal

ABSTRACT

The analysis of human location histories is currently getting an increasing attention, due to the widespread usage of geopositioning technologies such as the GPS, and also of online location-based services that allow users to share this information. Tasks such as the prediction of human movement can be addressed through the usage of these data, in turn offering support for more advanced applications, such as adaptive mobile services with proactive context-based functions. This paper presents an hybrid method for predicting human mobility on the basis of Hidden Markov Models (HMMs). The proposed approach clusters location histories according to their characteristics, and latter trains an HMM for each cluster. The usage of HMMs allows us to account with location characteristics as unobservable parameters, and also to account with the effects of each individual's previous actions. We report on a series of experiments with a real-world location history dataset from the GeoLife project, showing that a prediction accuracy of 13.85% can be achieved when considering regions of roughly 1280 squared meters.

Author Keywords

Hidden Markov Models, Hierarchical Triangular Meshes, Location Prediction, GPS Trajectory Analysis.

ACM Classification Keywords

H.2.8 Database Applications: Data Mining.

General Terms

Experimentation, Human Factors

INTRODUCTION

The widespread usage of localization systems, such as the Global Positioning System (GPS), are making it possible to collect interesting data in many different domains. Modern geopositioning technologies have become ubiquous, and there are nowadays many possibilities for effectively tracking the position of individuals over time. Simultaneously, location-based social networks such as FourSquare[1], or GPS track sharing services such as GoBreadCrumbs[2], are nowadays being increasingly used as a means to store and share human location histories.

One way to use these data is to interpret the shared location histories as the observed portion of complex sequential systems which include hidden contextual variables, such as the activities and goals motivating individual's movements at each time period, in the traces of visits to particular locations. If a distribution over the possible values in such a system can be estimated, then we can use previously observed paths to make inference about hidden states, and afterwards to make informed guesses about other places likely to follow the observed part of these paths.

This paper addresses the development and evaluation of generative models that, by capturing the sequential relations between places visited in a given time period by particular individuals, support the analysis and inference of statistical patterns for predicting future locations to be visited. Specifically, we propose an hybrid method based on Hidden Markov Models, in which human location histories are first clustered according to their characteristics, and in which the clusters are then used to train different Hidden Markov Models, one for each cluster. By leveraging on HMMs for modeling the sequences of visits, we have that the proposed method can account with location characteristics as unobservable parameters, and also with the effects of each individual's previous actions. Through experiments with a real-world location history dataset from the GeoLife project, we measured the prediction accuracy of the proposed method under different configurations. The overall obtained results correspond to a prediction accuracy of 13.85%, when considering regions of roughly 1280 squared meters. In terms of the geospatial distance between the true location of the user, and the geospatial coordinates that are predicted, the best results correspond to an average distance of 143.506 Kilometers, and a median distance of 4.957 Kilometers.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

[1] https://foursquare.com/
[2] http://www.gobreadcrumbs.com/

The rest of this paper is organized as follows: Section 2 presents the most important related work. Section 3 details the proposed method, discussing the usage of a technique known as the hierarchical triangular mesh for representing individual locations, detailing the clustering of location histories, and presenting the training and inference of patterns with Hidden Markov Models. Section 4 presents the experimental validation of the proposed method, describing the evaluation methodology and discussing the obtained results. Finally, Section 5 summarizes our conclusions and presents possible directions for future work.

RELATED WORK

Many previous works have addressed the issue of computing with spatial trajectories, and a detailed survey is given in the book by Zheng and Zhou [26]. Moreover, several previous works have specifically focused on the analysis of human location histories, concluding that human trajectories show a high degree of temporal and spatial regularity, following simple and reproducible patterns [3, 8]. In brief we have that previously proposed methods for the analysis of location histories can be classified, according to the manner by which data are modeled, into three general distinct approaches, namely (i) state-space models, (ii) data mining techniques, and (iii) template matching techniques.

State-space models attempt to capture the variation in spatial sequences through sequence models such as generative Hidden Markov Models (HMMs) [17], discriminative Conditional Random Fields (CRFs) [21, 19], or extensions of these two well-known approaches [4, 15]. Generally, these models have been used successfully in dealing with uncertainty (i.e., they generalize well), but they also suffer from high training complexity. In the case of location prediction, generative approaches such as HMMs can naturally be used, since they support the generation of possible future visits and the estimation of an associated probability.

Data mining techniques, on the other hand, explore frequent patterns and association rules, by defining a trajectory as an ordered sequence of time-stamped locations, and using sequence analysis methods such as modified versions of the Apriori algorithm [13, 14]. Most previous data mining methods attempt to maximize confidence with basis on previous occurrences (i.e., they do not generalize as well as state-space models), and they often do not consider any notion of spatial and/or temporal distance.

The third type of approaches, which are based on template matching, compare extracted features to pre-stored patterns or templates, using similarity metrics specific for sequential or time-series data. These similarity metrics include dynamic time warping and other sequence alignment approaches that are essentially variations of the edit distance computed between the sequences (e.g., edit distance with real penalty or edit distance on real sequence). They also include algorithms based on the longest common subsequence, or even other heuristic algorithms similar to those used in more traditional string matching problems [18, 16]. Template matching techniques have also been reported to have issues with high runtime complexity, noise intolerance, or in dealing with spatial activity variation.

Asahara et al. proposed a state-space modeling method for predicting pedestrian movement on the basis of a mixed Markov-chain model (MMM), taking into account a pedestrian's personality as an unobservable parameter, together with the pedestrian's previous status [1]. The authors report an accuracy of 74.4% for the MMM method and, in a comparison over the same dataset, the authors reported that methods based on Markov-chain models, or based on Hidden Markov Models, achieve lower prediction rates of about 45% and 2%, respectively for each of these two cases.

Sébastien et al. extended a previously proposed mobility model called the Mobility Markov Chain (MMC), in order to consider the n previous visited locations [7]. This proposal essentially corresponds to a higher order Markov model. Experiments on different datasets showed that the accuracy of the predictions ranges between 70% to 95%, but they also show that improvements obtained by increasing $n > 2$ do not compensate for the computational overhead.

Morzy, in turn, proposed data mining methods for predicting the future location of moving objects [14]. He extracts association rules from the moving object database and, given a previously unseen trajectory of a moving object, he uses matching functions to select the best association rule that matches the trajectory, afterwards relying on this rule for the prediction. This author reports on an accuracy of 80% for the best configuration of the proposed system.

Other authors still have proposed hybrid sequence analysis approaches, combining multiple types of information. For instance Jeung et al. proposed a hybrid prediction approach which estimates future locations based on pattern information extracted from similar trajectories, together with motion functions based on recent movements [9]. Lu and Tseng used a sequence similarity measure to evaluate the similarity between location histories, afterwards using a clustering algorithm to form a user cluster model of the location histories, based on the similarity measure. Then, using the clusters, the authors predict the movement of individuals, i.e., their next location [12]. Ying et al. proposed an approach for predicting the next location of an individual based on both geographic and semantic features of trajectories, using a cluster-based prediction strategy which evaluates the next locations with basis on the frequent behaviors of similar users in the same cluster, as determined by analyzing common behavior in terms of the semantic trajectories [22].

Previous works have also attempted to analyze human location histories through non-sequential unsupervised approaches based on probabilistic topic models, such as the topic model known as Latent Dirichlet Allocation (LDA) [6]. Probabilistic generative models have been typically used for analyzing document collections, by identifying the latent structure that underlies a set of observations (i.e., words contained within documents). In the case of location histories, the idea behind these models is to represent trajectories as a mixture

of topics, which in turn are modeled as probabilistic distributions over the possible locations.

In this paper, we study the problem of modeling human location histories for predicting the next places to be visited, using a very simple approach based on Hidden Markov Models, and evaluating its limitations on real-world data.

THE PROPOSED METHOD

This paper proposes a simple method for predicting the future locations of mobile individuals, on the basis of their previous visits to other locations, and leveraging on Hidden Markov Models for capturing the patterns embedded in previously collected location histories.

In the proposed method, the previously collected location histories are first clustered according to their characteristics (i.e., according to the temporal period in which they occurred). Afterwards, the clusters are used to train different Hidden Markov Models (HMMs) corresponding to the different types of location histories (i.e., one HMM for each cluster). Given a new sequence of visits, from which we want to discover the particular location that is more likely to be visited next, we start by finding the cluster that is more likely to be associated to the particular sequence of visits being considered in the prediction task, and then we use inference over the corresponding HMM in order to discover the most probable following location.

Each of the places in a given sequence of visits is associated to a timestamp and to the corresponding geospatial coordinates of latitude and longitude. The initial clustering of sequences is based on the temporal period associated to each sequence, and we use the timestamp associated to the last place visited in the sequence, in order to group sequences according to three clusters, namely (i) a cluster for sequences whose visits are made on weekdays by daytime (i.e., from 7AM to 7PM), (ii) a cluster for sequences whose visits are made on weekdays by nighttime (i.e., from 7PM to 7AM in the next day), and (iii) a cluster for sequences containing places visited in weekends. Notice that the proposed clustering approach is only a very simple approximation that uses the timestamp of the last visited location, and not the covered temporal periods themselves. More sophisticated clustering algorithms could indeed be used in the first step of the proposed approach (e.g., clustering through the usage of a mixture of Gaussians), but we leave this for future work.

The Hidden Markov Models for each cluster use only information regarding the geospatial positions (i.e., information derived from the geospatial coordinates of latitude and longitude) for each visited place. Prior to using HMM models, each of the places in a given sequence is first pre-processed in order to convert the real continuous values associated to the geospatial coordinates of latitude and longitude, into discrete codes associated to specific regions. This is done so that the learning of HMMs can be done more efficiently, using categorical distributions over the possible regions in the HMM states, instead of using more complex distributions (e.g., multivariate Gaussian distributions or Von Mises-Fisher spherical distributions) over the real values associated to the geospatial coordinates. To do this discretization, we used the hierarchical triangular mesh[3] approach to divide the Earth's surface into a set of triangular regions, each roughly occupying an equal area of the Earth [20, 5].

In brief, we have that the Hierarchical Triangular Mesh (HTM) offers a multi-level recursive decomposition of a spherical approximation to the Earth's surface. It starts at level zero with an octahedron and, by projecting the edges of the octahedron onto the sphere, it creates 8 spherical triangles, 4 on the Northern and 4 on the Southern hemisphere. Four of these triangles share a vertex at the pole and the sides opposite to the pole form the equator. Each of the 8 spherical triangles can be split into four smaller triangles by introducing new vertices at the midpoints of each side, and adding a great circle arc segment to connect the new vertices with the existing ones. This sub-division process can repeat recursively, until we reach the desired level of resolution. The triangles in this mesh are the regions used in our representation of the Earth, and every triangle, at any resolution, is represented by a single numeric ID. For each location given by a pair of geospatial coordinates on the surface of the Earth, there is an ID representing the triangle, at a particular resolution, that contains the corresponding point. Notice that the proposed representation scheme contains a parameter k that controls the resolution, i.e., the area of the triangular regions. In our experiments, we made tests with the values of 18 and 22 for the parameter k, which correspond to triangles of roughly 1280 squared meters, and of 5.002 squared meters, respectively. With a resolution of k, the number of regions n used to represent the Earth corresponds to $n = 8 \times 4^k$. Figures 1 and 2 illustrate the decomposition of the Earth's sphere into a triangular mesh.

Figure 1. Decomposition of the Earth's surface for triangular meshes with resolutions of zero, one, and two.

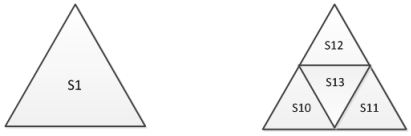

Figure 2. Decomposition of circular triangles.

HMM Training and Inference

Hidden Markov Models (HMMs) are a well-known approach for the analysis of sequential data, in which the sequences are assumed to be generated by a Markov process with unobserved (i.e., hidden) states [17]. Thus, by using HMMs for

[3] www.skyserver.org/htm/Old_default.aspx

modeling location sequences, the states governing the moving agent's decisions are not directly visible, but the visited locations, dependent on the state, are visible. Each state has a probability distribution over the possible locations to be visited, in our case a categorical distribution over the triangles representing each possible area in the Earth's surface. Each state has also a probability distribution over the possible transitions to the state that is going to govern the decision about the next place to be visited in the sequence.

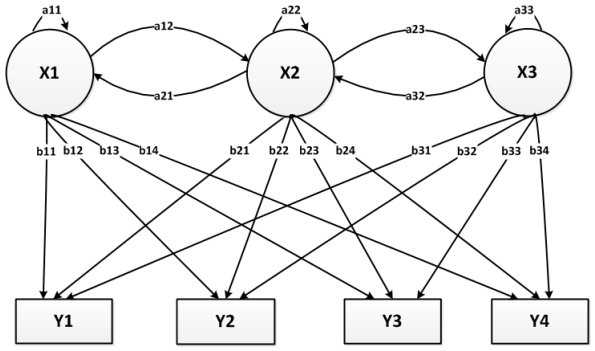

Figure 3. Example Hidden Markov Model.

The diagram in Figure 3 shows the general architecture of an instantiated HMM. Each shape in the diagram represents a random variable that can adopt any of a number of values. The random variable $x(t)$ is the hidden state at time t (in the model from the above diagram, $x(t) \in \{x1, x2, x3\}$). The random variable $y(t)$ is the location visited at time t (with $y(t) \in \{y1, y2, y3, y4\}$). The arrows in the diagram denote conditional dependencies. From the diagram, it is clear that the conditional probability distribution of the hidden variable $x(t)$ at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable $x(t-1)$, and thus the values at time $t-2$ and before have no influence. This is called the Markov property. Similarly, the value of the observed location $y(t)$ only depends on the value of the hidden variable $x(t)$, at time t.

Several inference problems can be addressed on top of instantiated HMMs, one of them being the computation of the probability for a given observation sequence (i.e., a sequence of visits to locations). The task is to compute, given the parameters of the instantiated model, the probability of a particular output sequence being observed. This requires computing a summation over all possible state sequences. The probability of observing a particular sequence in the form $Y =< y(1), y(2), \ldots, y(L) >$, of length L, is given by:

$$P(Y) = \sum_{X} P(Y \mid X)P(X) \tag{1}$$

In the formula, the sum runs over all possible hidden-node sequences $X =< x(1), x(2), \ldots, x(L) >$. Applying the principle of dynamic programming, this problem can be handled efficiently, using a procedure known as the the forward algorithm, which we outline further ahead.

As for HMM parameter learning, the task is to find, given an output sequence X or a set of such sequences, the best set of state transition and output probabilities, i.e., the values for $a_{i,j}$ and $b_i(k)$ from Figure 3. The task is usually to derive the maximum likelihood estimate of the parameters of the HMM, given the set of output sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the Baum-Welch algorithm, which is an example of a forward-backward algorithm and a special case of the well-known expectation-maximization (EM) algorithm.

For the following presentation of the Baum-Welch algorithm, we can describe an HMM by $\lambda = (A, B, \pi)$, where A is a time-independent stochastic transition matrix between states, B is a stochastic matrix with the probabilities of outputting a particular observation in a given state, and π is the initial state distribution (i.e., a matrix enconding transition probabilities for the particular case of the first position in the sequences). Given a single observation sequence corresponding to $Y =< y(1); y(2); \ldots; y(L) >$, the Baum-Welch algorithm finds $\lambda^* = \max_\lambda P(Y|\lambda)$, i.e. the HMM λ that maximizes the probability of the observation sequence Y.

The Baum-Welch initialization sets $\lambda = (A, B, \pi)$ with random initial conditions. The algorithm then updates the parameters of λ iteratively until convergence, or until a given number of steps has been reached.

Is an expectation step, the forward algorithm is first used to define $\alpha_i(t) = P(y(1) = o_1, \ldots, y(t) = o_t, X(t) = i|\lambda)$, which is the probability of seeing the partial observable sequence $o_1, \ldots, o_t$ and ending up in state i at time t. We can efficiently calculate $\alpha_i(t)$ recursively through dynamic programming, applying the following two equations:

$$\alpha_i(1) = \pi_i b_i(o_1) \tag{2}$$

$$\alpha_i(t+1) = b_i(o_{t+1}) \sum_{j=1}^{N} \alpha_j(t) \times a_{j,i} \tag{3}$$

In the above formula N refers to the size of the set of possible states, a_{ij} refers to transition probabilities, and $b_j(o)$ refers to the emission probabilities. A backwards procedure is also used to compute the probability of the ending partial sequence $o_{t+1}, \ldots, ot$, given that we started at state i at position L. Similarly to the previous case, we can compute a value $\beta_i(t)$ recursively, through:

$$\beta_i(L) = 1 \tag{4}$$

$$\beta_i(t) = \sum_{j=1}^{N} \beta_j(t+1) a_{i,j} b_j(o_{t+1}) \tag{5}$$

Using α and β we can calculate the following variables:

$$\gamma_i(t) \equiv p(Y(t) = i|X, \lambda) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{6}$$

$$\begin{aligned} \xi_{i,j}(t) &\equiv p(Y(t) = i, Y(t+1) = j|X, \lambda) \\ &= \frac{\alpha_i(t) a_{i,j} \beta_j(t+1) b_j(o_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{i,j} \beta_j(t+1) b_j(o_{t+1})} \end{aligned} \tag{7}$$

Having γ and ξ, in the maximization step, one can define update rules for the HMM as follows:

$$\overline{\pi_i} = \gamma_i(1) \tag{8}$$

$$\overline{a_{i,j}} = \frac{\sum_{t=1}^{L-1} \xi_{i,j}(t)}{\sum_{t=1}^{L-1} \gamma_i(t)} \tag{9}$$

$$\overline{b_i}(k) = \frac{\sum_{t=1}^{L} \delta_{o_t,o_k} \gamma_i(t)}{\sum_{t=1}^{L} \gamma_i(t)} \tag{10}$$

Note that the summation in the nominator of $\overline{b_i}(k)$ is only made over observed symbols equal to o_k, i.e., $\delta(o_t, o_k) = 1$ if $o_t = o_k$, and zero otherwise. Using the updated values of A, B and π, a new iteration of the above procedure is preformed, and we repeat this until convergence.

For more information about the algorithms typically used in hidden Markov modeling problems, please refer to the tutorial by Rabiner [17].

In the particular application addressed in this paper, we used the Baum-Welch approach for estimating the parameters of our Hidden Markov Models, and the task of human movement prediction is reduced to the particular HMM inference problem of computing, given a set of sequences of the form $Y =< y(0), y(1), \ldots, y(L), y(L_{next}) >$, in which the first L positions correspond to places already visited by a particular pedestrian, and in which position L_{next} encodes a possible location to be visited next, the sequence that has the highest probability and, correspondingly, the place L_{next} that is more likely to be visited next. From a given sequence of previous visits, we (i) compute the set of sequences corresponding to all possible next places to be visited, (ii) use the forward algorithm to compute the probability of all such sequences, and (iii) return the next place corresponding to the sequence with the highest probability.

The general approach that is proposed for addressing the location prediction task could also be made to rely on more sophisticated models, such as high-order HMMs of even the recently proposed infinite-HMM model [2]. However, in this paper, we only report on experiments with regular first-order HMMs, and leave other options to future work.

EXPERIMENTAL EVALUATION

We implemented the proposed approach for pedestrian movement prediction through an existing implementation of the algorithms associated with Hidden Markov Models (i.e., the Baum-Welch and the forward-backward procedures), latter validating the proposed ideas through experiments with the GeoLife dataset, i.e. a GPS trajectory dataset collected in the context of the GeoLife project from Microsoft Research Asia, by 178 users in a period of over three years from April 2007 to Oct. 2011 [24, 23, 25]. In the GeoLife dataset, a GPS trajectory is represented by a sequence of time-stamped points, each of which containing latitude and longitude coordinates. The full dataset contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. The trajectories were recorded by different GPS loggers and GPS-phones, and have a variety of sampling rates, with 91% percent of the trajectories being logged in a dense representation, e.g. every 15 seconds or every 10 meters per sample. The authors claim that the GeoLife dataset recorded a broad range of user's outdoor movements, including not only life routines like going home and going to work, but also some entertainment and sports activities, such as shopping, sightseeing, dining, or hiking. This fact makes the dataset ideal for the purpose of validating pedestrian movement prediction methods.

We experimented with different configurations of the propose method, namely by (i) varying the number of states in the Hidden Markov Models between 10, 15 and 20 states, (ii) varying the resolution of the triangular decomposition of the Earth between resolutions of 18 and 22, and (iii) using the clustering method prior to the training of Hidden Markov Models, versus using a single HMM.

When using the GeoLife dataset, we converted the latitude and longitude coordinates to triangular regions, according to the approach based on the hierarchical triangular mesh, and then we removed from the considered sequences all elements $x(t+1)$ that corresponded to the same triangular region given in position $x(t)$ (i.e., we removed all duplicate consecutive locations from each trajectory in the dataset). All the considered trajectories had a minimum length of 10 locations, and a maximum length of 25 locations (i.e., we only kept the last 25 different locations that were visited). In order to use an equal number of sequences in the training of our different Hidden Markov Models (i.e., in each of the clusters) we randomly selected 3465 different sequences for each of the three clusters. This value corresponds to the number of sequences in the smallest cluster, when considering the entire dataset. Similarly, 3465 different sequences were selected for the case of the experiments with the single HMM. The considered evaluation procedure was based on removing the last location in each of the validation trajectories, afterwards generating predictions for the next place to be visited, and finally checking the prediction against the true location that a given pedestrian visited next.

The training of the HMM models took between 21 and 318 minutes on a standard laptop PC, with results varying according to the number of states.

Table 1 shows the number of different visited locations considered during the training process. In the case of the multiple HMMs, and since each cluster used a different set of sequences, the values that are shown correspond to the average of the values for the different clusters.

Table 2 shows the obtained results for the different configurations, measuring the quality of the obtained results through the Precision@1, Precision@5, and the Mean Reciprocal Rank (MRR) evaluation metrics. Precision@1 measures the percentage of times in which we found the correct next location, whereas Precision@5 measures the percentage of times in which the correct next location was given in the top-five most probable locations. The mean reciprocal rank metric

	Geospatial Resolution	
	18	22
Multiple HMMs	26066	66063
Single HMM	24807	59778

Table 1. Number of different visited sites.

measures the average of the reciprocal ranking positions for the correct next location, in each of the ranked lists with all possible locations that are produced for each trajectory.

Table 2 also illustrates the quality of the obtained results through the average geospatial distance, computed through Vincenty's formulae[4], between the true coordinates of latitude and longitude associated to next place visited by the user, and the centroid coordinates for the triangular region, corresponding to the next location that is predicted. These distance values were measured both for the cases in which the predicted location was correct (i.e., the predicted triangular region contained the next geospatial coordinates of latitude and longitude), and for the cases in which the predicted location was incorrect.

Table 3 shows, instead, the best results obtained with the multiple HMMs corresponding to the different clusters, presenting the results obtained for each of the different clusters.

Finally, in Figure 4 we plot the cumulative distance errors obtained with the best configurations, for the cases in which we used the single or the multiple HMMs, i.e., the distances between the expected locations and the predicted locations, in each sequence. The red line corresponds to the errors for the case of the multiple HMMs, while the black line corresponds to distance errors in the case of the single HMM the lines shown in the chart are not directly comparable, since there are 3 times more sequences in the case of the red line. Still, one can notice that most of the cases correspond to small errors in terms of distance, and indeed we have that the median value for the best configuration corresponds to 4.957 Kilometers.

Overall, the approach were a single HMM was used achieved better accuracies. The single HMM with the configuration using a geospatial resolution of 18, and considering 15 hidden states, achieved the best results in our experiments. The errors in terms of the geospatial distance are generally smaller when we consider smaller values of geospatial resolution (i.e., large prediction areas), and also when we consider higher values for the number of states in the HMMs. Notice that the area of the triangular regions with a resolution of 22 is much smaller than the area of the triangular regions with a resolution of 18. Hence, the accuracies from the experiments using a resolution of 22 are smaller than the accuracies obtained with the resolution of 18, leading also to a worse overall performance in terms of the geospatial distance.

CONCLUSIONS

[4] http://en.wikipedia.org/wiki/Vincenty's_formulae

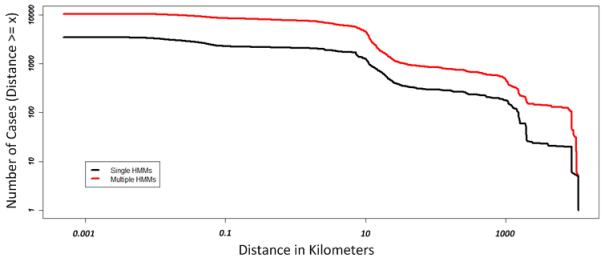

Figure 4. Distribution of the distance errors.

This paper presented an hybrid method for predicting individual's movements on the basis of Hidden Markov Models. The proposed approach clusters human location histories according to their characteristics (i.e., according to the temporal period in which the visits where made), and latter trains a Hidden Markov Model for each cluster, this way accounting with location characteristics as unobservable parameters, and also accounting with the effects of each pedestrian's previous actions. We report on a set of experiments with different configurations of the proposed method, and using a real-world location history dataset from the GeoLife project. We measured a prediction accuracy of about 13.85% with the best performing method. In terms of the geospatial distance between the true location of the user, and the geospatial coordinates that are predicted, the best results correspond to an average distance of 143.506 Kilometers, and a median distance of 4.957 Kilometers.

Despite the interesting results, there are also many ideas for future work. A particular idea that we would like to try relates to improving the training of Hidden Markov Models through posterior regularization, a method that allows one to incorporate indirect supervision (e.g., the fact that consecutive locations that are located close-by should be more probable to occur than others) via constraints on posterior distributions of probabilistic models with latent variables [6].

Many previous works have also addressed the subject of analyzing and classifying sequential data collected from multiple domains, using methods such as sliding window classifiers, Conditional Random Fields (CRFs), Averaged Perceptrons (APs), Structured Support Vector Machines (SVM struct), Max Margin Markov Networks (M3Ns), Search-based Structured Prediction (SEARN) models, or Structured Learning Ensemble models (SLEs). Authors like Thomas G. Dietterich, or Nguyen and Guo, have provided good surveys in the area [15, 4]. For future work, we would also like to experiment with discriminative models, such as those referenced above, in order to address related sequence analysis methods, such as the classification of human location histories according to semantic categories [11, 10].

Acknowledgements

The authors would like to express their gratitude to Fundação para a Ciência e a Tecnologia (FCT), for the financial support offered through the project grant corresponding to

	Parameters		Accuracy			Average Distance in Km		
	Nr. States	Geo. Resolution	P@1	P@5	MRR	Correct	Incorrect	Average
Multiple HMMs	10	18	05.89	14.55	0.105	0.0143	197.855	186.189
	15	18	05.56	**15.69**	0.109	0.0145	197.164	186.201
	20	18	**06.02**	15.44	**0.110**	0.0142	198.197	186.262
	10	22	00.12	00.56	0.005	**0.0008**	193.399	193.157
	15	22	00.17	00.72	0.006	0.0009	**186.098**	185.776
	20	22	00.14	00.72	0.006	0.0008	186.382	186.113
Single HMM	10	18	07.09	24.79	0.149	0.0125	154.797	143.808
	15	18	**13.85**	**26.40**	**0.201**	0.0118	166.580	**143.506**
	20	18	09.26	25.59	0.169	0.0129	158.490	143.808
	10	22	00.08	00.20	0.004	**0.0008**	149.750	149.620
	15	22	00.14	00.89	0.008	0.0010	148.949	148.734
	20	22	00.31	00.75	0.008	0.0010	**148.636**	148.164

Table 2. Results for different configurations of the proposed location prediction method.

	Accuracy			Average Distance in Km		
Cluster	P@1	P@5	MRR	Correct	Incorrect	Average
Weekday 7AM to 7PM (Morning)	08.16	15.55	0.122	0.0142	195.582	179.609
Weekday 7PM to 7AM (Evening)	04.90	16.47	0.110	0.0150	216.451	205.832
Weekend	04.99	14.31	0.097	0.0133	182.455	173.346

Table 3. Results with the multiple HMMs corresponding to different clusters, considering 20 states and a geospatial resolution of 18.

reference PTDC/EIA-EIA/109840/2009 (SInteliGIS).

REFERENCES

1. A. Asahara, K. Maruyama, A. Sato, and K. Seto. Pedestrian-movement prediction based on mixed Markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 25–33. ACM, 2011.
2. M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The Infinite Hidden Markov Model. In *Proceedings of the 16th Conference on Neural Information Processing Systems*, pages 29–245. MIT Press, 2002.
3. Y. Chon, H. Shin, E. Talipov, and H. Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Proceedings of the 10th IEEE International Conference on Pervasive Computing and Communications*, pages 206–212. IEEE Computer Society, 2012.
4. T. G. Dietterich. Machine Learning for Sequential Data: A Review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30. Springer-Verlag, 2002.
5. G. Dutton. Encoding and Handling Geospatial Data with Hierarchical Triangular Meshes. In *Proceedings of the 7th Symposium on Spatial Data Handling*, pages 505–518. Spatial Effects, 1996.
6. K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. In *ACM Transaction Intelligent System Technology*, pages 1–27. ACM, 2011.
7. S. Gambs, M. Killijian, D. P. Cortez, and N. Miguel. Next place prediction using mobility Markov chains. In *Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility*, pages 1–6. ACM, 2012.
8. M. C. Gonzalez, C. A. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, (7196):779–782, 2008.
9. H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A Hybrid Prediction Model for Moving Objects. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pages 70–79. IEEE Computer Society, 2008.
10. J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. The Mobile Data Challenge:Big Data for Mobile Computing Research. In *Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing*, pages 1–8. ACM, 2012.
11. L. Liao, D. Fox, and H. Kautz. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *International Journal of Robotics Research.*, (1):119–134, 2007.
12. E. H. Lu and V. S. Tseng. Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In *Proceedings of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware*, pages 273–278. IEEE Computer Society, 2009.
13. A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. WhereNext: a location predictor on trajectory pattern

mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 637–646. ACM, 2009.

14. M. Morzy. Mining Frequent Trajectories of Moving Objects for Location Prediction. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 667–680. Springer-Verlag, 2007.
15. N. Nguyen and Y. Guo. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 681–688. ACM, 2007.
16. O. Ossama and H. M. O. Mokhtar. Similarity Search in Moving Object Trajectories. In *Proceedings of the 15th International Conference on Management of Data*, pages 1–6. Computer Society of India, 2009.
17. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Readings in Speech Recognition*, pages 267–296. Morgan Kaufmann Publishers, 1990.
18. D. E. Riedel, S. Venkatesh, and W. Liu. Recognising online spatial activities using a bioinformatics inspired sequence alignment approach. *Pattern Recognition*, (11):3481–3492, 2008.
19. C. Sutton and A. McCallum. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, To be published.
20. A. S. Szalay, J. Gray, G. Fekete, P. Z. Kunszt, P. Kukol, and A. Thakar. Indexing the Sphere with the Hierarchical Triangular Mesh. *Computing Research Repository*, pages 58–65, 2007.
21. D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1–8. ACM, 2007.
22. J. J. Ying, W. Lee, T. Weng, and V. S. Tseng. Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 34–43. ACM, 2011.
23. Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Ma. Understanding mobility based on GPS data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 312–321. ACM, 2008.
24. Y. Zheng, X. Xie, and W. Ma. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Data Engineering Bulletin*, (2):32–39, 2010.
25. Y. Zheng, L. Zhang, X. Xie, and W. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, pages 791–800. ACM, 2009.
26. Y. Zheng and X. Zhou. *Computing with Spatial Trajectories*. Springer, 2011.