Trip destination prediction based on past GPS log using a Hidden Markov Model

J. A. Alvarez-Garcia[†], J. A. Ortega[†], L. Gonzalez-Abril[‡], and F. Velasco[‡]

[†] Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain

[‡] Applied Economics I Dept., University of Seville, 41018 Seville, Spain

Abstract

In this paper, a system based on the generation of a Hidden Markov Model from the past GPS log and current location is presented to predict a user's destination when beginning a new trip. This approach drastically reduces the number of points supplied by the GPS device and it permits a "support-map" to be generated in which the main characteristics of the trips for each user are taken into account. Hence, in contrast with other similar approaches, total independence from a street-map database is achieved.

Key words: Knowledge discovery; Machine learning; Predictive HMM; Information retrieval

1 Introduction

In developed societies, life is lived in a sedentary and comfortable environment where the displacements of a person within a geographical region usually follow a pattern. Most of these displacements are carried out towards known places such as home, the workplace, relatives' houses, a favourite cinema or fashionable shopping centre. Hence, patterns of these movements can be studied and a prediction system can be designed by allowing all the places where a user could go to be stored in order to anticipate arrival.

It is worth noting that from the elimination of selective availability (that introduced random errors of up to a hundred metres) on May 1, 2000, GPS

receptors have evolved from research tools to become consumer goods. The E-911 mandate[1] in the United States and the E112 recommendation in Europe have boosted the development and use of systems based on GPS. Nowadays, millions of civilian users have a GPS navigation system to help them drive to previously unknown locations. However, future location and trip prediction systems are on-going topics in research, and not a software product in the same way as navigation systems. Along these lines, some interesting scenarios have been studied: predictions of meetings between two users and intelligent interruption [1], location to-do list [9], and optimization of a hybrid vehicle charge and discharge schedule [4].

The probabilistic models of destination prediction using GPS was born in 2003 with the work of Ashbrook and Starner [1]. A Markov model was used to find the next most likely destination based on those that had recently been visited from a set of previously clustered candidate destinations. In the same year, a Bayesian model of a traveller moving through an urban environment was presented in [10], and a year later this development was improved in [8] with a predictive hierarchical model that can learn and infer a user's daily movements using different means of transport. In 2006, the "Predestination" model [7] used Bayesian inference to obtain a probabilistic map of destinations; and a Hidden Markov Model (HMM) is given in [12] to predict the next link road and the final destination of an user. By following similar reasoning, an n^{th}-order Markov model is trained in [6] to probabilistically predict future road segments based on a short sequence of just-driven road segments.

It is important to note that all of the aforementioned work needs external information from the geographical environment. Hence, a street-map is required in [10,8,6,12] since all locations must be within its range. The street-map is modelled as a directed graph whose edges correspond to streets or footpaths and whose vertices correspond to intersections, hence by knowing the current location, the next transitions can be predicted. Furthermore, to adapt the GPS points received from the receptor to the map database, a process called map-matching is developed in these papers. This process is set up to work on a road level but not on a point level, since all approaches previous to this paper are based on partial trips which work with all the roads that the user traverses until the final destination.

Moreover, the street-map database needed in these earlier probabilistic models makes their predictions local to the cities on which they are dependent. All these papers have shown that a common user (not including salespeople, delivery people, etc.) only traverses a reduced percentage of all the existing roads in a city, and hence, in our approach, a Geographical Information System (GIS) is not a constraint of the construction of the user's geographical

[1] http://www.fcc.gov/911/enhanced/

model.

In this paper, the technique used to extract probable destinations is similar to that given in [10,8,6,12], that is, to analyze the GPS journeys of users by extracting clustered destination points from these journeys. Furthermore, a Hidden Markov Model is generated in our approach by using a new procedure to extract the significant points which have been used to predict the route and finally the destination of one user when the trip is not yet completed.

The generated HMM [2] is set up to detect an invisible state process by using a visible observation sequence of another process. In this approach, the invisible process is the goal to reach (the destination place) and the visible observations are the sequence of significant points, called support points, that compose a route. Therefore, map-matching problems are avoided in the developed system since it is independent of any street-map.

The paper is structured as follows. In Section 2, the complete process to obtain the data from the GPS logs of different users is described. The HMM for predicting destinations is presented in Section 3. A practical implementation based on this model using a corpus of real driving, running, walking and biking data is provided in Section 4. Conclusions are drawn in the final section.

2 Retrieving information from users' GPS logs

The process starting from the users' log retrieval to the extraction of the most relevant information is described in this section.

2.1 Obtaining and filtering data

For the sake of clarity, an explanation is given of how the 6 datasets were obtained in our experiment. The data was retrieved from GPS "Wintec 100" receiver devices [3] capable of measuring and storing up to 12,600 geo-positional points. The device had to be carried everywhere by the user and also charged, the trips downloaded to a computer through an USB connection, and then the downloaded information sent to a central database. These data are points with geographical and temporal information, covering a period of more than 7 days and at least 18 trips for each user. A portion of one set can be seen in Figure 3 a), where each point is labelled with the time and date of the trip.

[2] For further details on this topic, see [3,2,11].
[3] http://www.wintec.tw

It is well-known that when a GPS system is used, some outlier points exist caused by obscured line of vision, device cold starts, and other satellite disruption phenomena, and therefore datasets must be filtered. It is also worth noting that some users configured their devices with a data sampling interval of 1 second and others with 5 seconds. Hence, in order to obtain comparable datasets and to avoid redundant points, all of the datasets have been filtered by using the following rule [4]: "The minimum distance between two consecutive points must be at least 30 metres". It should be stressed that with this rule, the percentage of filtered data is usually high as can be seen in our experiment (see Table 1).

For the purpose of managing the journeys of each user, the sequence of points of each dataset is segmented into trips (a set of chronologically-ordered, time-stamped GPS data points where the first is the origin and the last is the destination). To this end, a widely used segment rule is chosen which looks for time gaps between two consecutive recorded points. Previous work selected a threshold ranging from 3 minutes [5] to 10 minutes [1]. This variety is produced by several elusive features, such as the amount of stopping time that users might consider significant, the maximum time waiting at a traffic light, or the stationary time in traffic jams. In our case the threshold is set to five minutes because it is an adequate time (it was observed that if longer than five minutes is chosen then two trips can be linked in only one trip), and since there is difficulty finding a traffic light or traffic jams where more than 5 minutes is spent in the same place. Trips with fewer than 10 points are also filtered out because indoor situations are empirically detected where the receptor catches very few valid GPS points.

2.2 Extracting knowledge

Once the filtered datasets are obtained, the most significant observations and frequent destinations are extracted from our model.

2.2.1 Frequent destinations

Due to the fact that users did not switch off their GPS's in exactly the same place when they reached a destination, there are a lot of close end points of trips near frequently-visited places. A clustering process is carried out in order to determine the probable destinations of each user. The final points of each trip were clustered. A threshold of 0.2 miles (320 m. approximately) is used in [1] but this measure depends on population density, city or town distribution

[4] Other similar rules have been considered but this rule has been chosen into account due to being the simplest and that the results are very promising.

and users' pattern of movement. In our work, this threshold is reduced in order to avoid clustering multiple places as if they were only one location. It is found that a threshold of 200 metres is a suitable measure for our 6 trip datasets.

Places with more than three visits are considered frequently-visited destinations. This number of visits is chosen since a greater minimum would lose data (one month of data from each user is the minimum provided) and fewer than three visits could prove problematic (some places were observed where the users had lost the GPS signal twice for long periods).

It should be pointed out that only the end point of each trip and not the point of origin is used as a candidate for frequently-visited destinations. This is due to common problems with the GPS signal, that is, the time required by a GPS receiver to acquire satellite signals and navigation data, and the calculation of a position solution (called a fix) could take from 10 seconds to some minutes, and hence the initial point yields less accurate data than the destination point.

This analysis is also carried out by all the aforementioned work [6,8,12,10] although these authors fail to take advantage of the information in order to extract the local "street-graph" for each user. This advantage is described in the next section.

2.2.2 From trips to a sequence of significant observations: support points

Each trip can have hundreds of points, and therefore if the inclusion of all trips in a statistical model is desired then it is necessary to reduce each trip to only a few manageable points. Furthermore, this reduction leads to another advantage: the computational time is reduced. On the other hand, the points chosen ("support points") must be significant, that is, these points must help us figure out where this user will finish the trip by simply knowing that this person is close to one of these points.

It should be pointed out that the consideration of "support points" is an important feature of our approach, and therefore an explanation is required here. When a user is placed before a crossroads with four possible paths to follow, the prediction is difficult, however, if, after a few seconds an observation point after this crossroads is chosen, then the destination prediction is much more accurate. Since no road map has been considered, "support points" are generated by analyzing the actual places where trips of each user, overlap, and not by the official crossroads of the place studied. Henceforth, the term "crossroads" will be used to denote any fork in the trajectory or an actual crossing of routes. A road map is not necessary since a local map is obtained from each particular user in our approach. Furthermore, the users frequently use the same path for their return trips but in the opposite direction, therefore

when this action occurs, two significant points at each observation point are considered which informs us of the direction followed by the user. Hence, the significant points, which are called "support points", are those placed after and before crossroads along the route of any pair of trips. Due to the fact that the same point could be traversed in different directions depending on the journey followed, the system labels each support point with the possible cardinal directions. A real example can be seen in Figure 2.

Let us illustrate the steps of generation of support points from Figure 1, which shows a set of 3 trips.

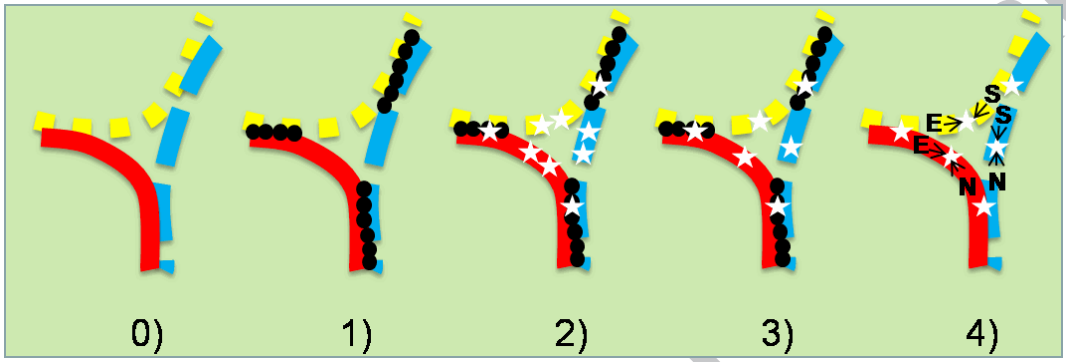

Figure 1. 0) shows 3 trips. In 1), black circles mark the common zones between each pair of trips, in 2) white stars point out the bounds of common zones, then in 3) these bounds are clustered, and finally in 4) each component of the cluster generates 2 support points which include the cardinal direction. For sake of simplicity, only the 6 support points generated by the 3 central stars are shown.

(1) Selection of all the common points for any pair of trips (see 1 in Figure 1).
(2) Analysis of the category of crossroads and selection of observation candidates. The candidates are the points on the bounds of the common point zone (see 2 in Figure 1 where the candidate points have been marked with stars).
(3) Clustering of nearby candidates. When numerous observation candidates are very close each other, only one point is chosen as support. (See 3 in Figure 1).
(4) Each candidate can be traversed in different directions, and hence one observation at the same point for each possible orientation (N, S, W, E) is considered. In 4 of Figure 1 two different orientations for each observation have been obtained.

An example of a real situation of support points is shown in Figure 2 where it can be seen that there are three trips (one is overlapped by another and therefore difficult to see) that generate a cross producing 8 support points, 2 for each location indicating the cardinal directions.

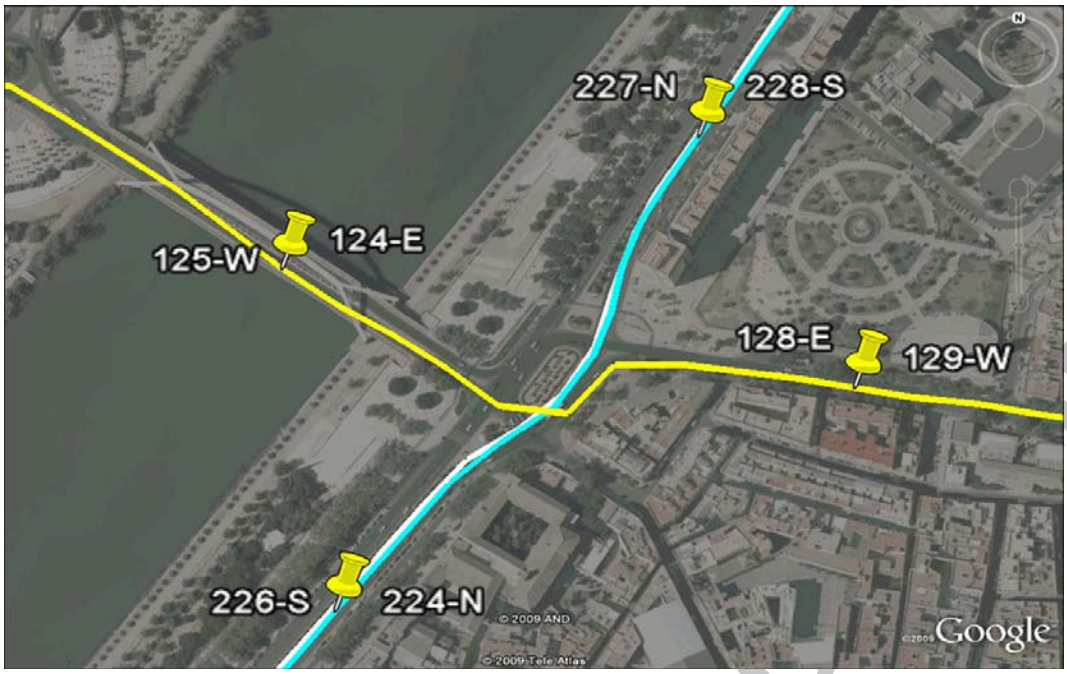

Figure 2. Support points generated in a real situation: Each support point is far enough from the crossroads to prevent the GPS device from indicating an erroneous significant observation. The same point can be traversed in different directions and hence the system labels each one with the possible directions of the trip.

It is worth noting that an important reduction of points in each trip is attained when this process is carried out. For example, in Figure 3 a portion of one of the studied datasets is shown before and after the "support points" are generated.

3 Hidden Markov Model Generation

Once the final destinations and support points are obtained, a Hidden Markov Model is designed to solve the trip destination prediction problem given only the data of a partial trip. The elements that formally define an HMM are (S, V, π, A, B) which must be adapted to our purpose:

- S is the set of N distinct states in the model. In this case, the states are the destination places of a user. It is worth noting that it is a logical supposition that any destination place will be the next place of origin, and hence it is not necessary to include origins in set S. The individual state is denoted by S_i and the state at observation i as q_i, that is, q_i denotes the final destination associated to the i-observation.
- V is the set of m distinct observation symbols per state (sub-states). The observation symbols correspond to the physical output of the system be-

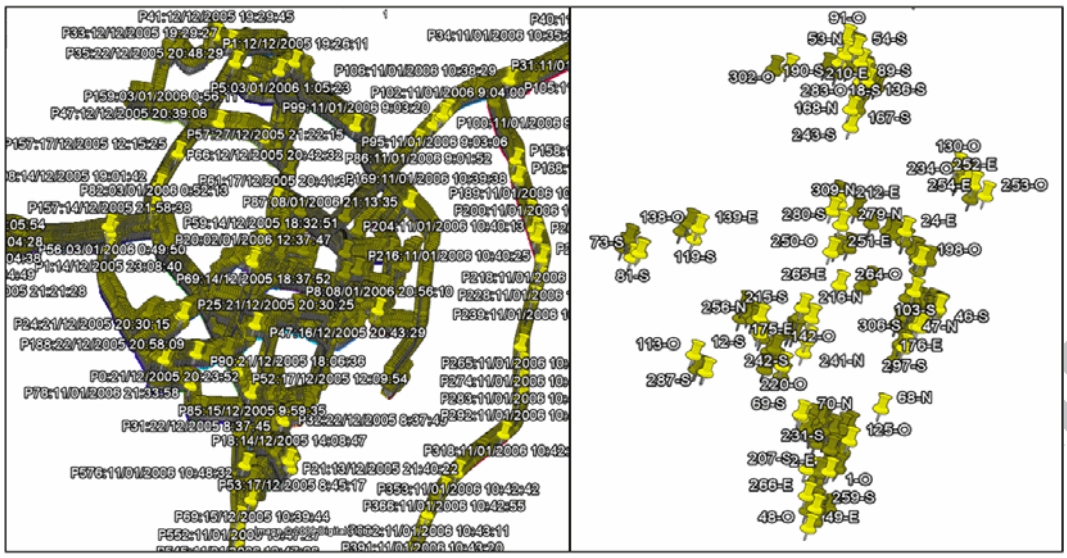

Figure 3. From trips to significant observations: in a) there is a partial set of points of one user (the label of each point represents the date and time); and in b) the support points generated after the complete process are shown.

ing modelled. For our model, these are the support points detailed in the previous section. Let us denote the support points as $V = \{v_1, \cdots, v_m\}$.

- The initial state distribution is $\pi = \{\pi_i\}_{i=1}^{N}$ where $\pi_i = P(q_1 = S_i)$. In our model, π_i denotes the probability that the final destination will be S_i, $i = 1, \cdots, N$ when a trip is beginning. It is considered that the initial state distribution is the probability of each destination in the training set to be the final destination. The most probable state is usually the home of the studied user.
- The state transition probability distribution $A = \{a_{ij}\}$ where $a_{ij} = P(q_{r+1} = S_j | q_r = S_i)$ for $1 \leq i, j \leq N$. This matrix contains the transition probabilities between each pair of states, that is, whether a user is close to a support point associated to state S_i, which is the probability that, at the next support point, the state will be S_j. In other words, a_{ij} shows the probability that while addressing one destination (S_i), our route changes and another destination (S_j) is addressed.
- The observation symbol probability distribution in state j, $B = \{b_j(k)\}$ where $b_j(k) = P(v_k \text{ at } r | q_r = S_j)$ for $1 \leq i \leq N$, and $1 \leq j \leq m$. Therefore, $bj(k)$ denotes the probability that the support point will be v_k, given that the final destination is S_j, that is, the probability of passing near a support point when destination S_j is addressed.

Once a state has been assigned to every observation (supervised learning) in the training set, the maximum likelihood estimators [11] are used to obtain an estimation of the (S, V, π, A, B) parameters of the model.

It is known that maximum likelihood estimators are vulnerable to overfitting if there are insufficient data. Indeed if there is a state S_k that is never used in the training set, then the estimation equations are undefined for this state. To prevent such a problem, it is preferable to add predetermined pseudocounts in the algorithm of calculation of the maximum likelihood estimators. Hence, the emission and transition pseudocounts help us prevent zero transition probability when using the Viterbi algorithm [13], and make our model more flexible.

4 Experiment

Firstly, the features of each dataset in the experiment are described. The number of studied users is six: two were selected from among the colleagues of our research group, another two from among our friends, and two students of the University of Seville (Spain). The main criterion of selection of these users was our prior knowledge about their different travelling habits. They showed distinct patterns of movements, the zone where each of them moves is diverse and trips could be by car, bicycle, running and/or walking. Another criterion was the diversity in the number of frequently visited places. Let us now see some particular characteristics of each user.

- User 1: The number of trips carried out by this user is 18, always by car and to only 3 destination places within the city of Seville.
- User 2: This user's data log reports information about 3 cities of Spain: Seville, Almeria and Huelva. The number of destinations is 7 and number of trips, 42. Trips are by car, by bicycle and 3 trips are on foot (running).
- User 3: 4 destination places within the city of Seville and a town 30 km from Seville. The number of trips is 60. Trips are by car and by bicycle.
- User 4: Only 4 destination places within the city of Seville and a town 10 km from Seville and 81 trips. All trips are by car. This is the most disciplined user because the GPS is never left switched on when not travelling and journeys are not by crowded roads so no traffic jams suffered. As a consequence, very little spurious data is generated.
- User 5: The number of destinations of this user is 12, the largest number of states of all users, found either in a town in Seville province or within the city of Seville. The number of trips is 213. Trips are by car and 4 on foot (walking).
- User 6: The largest number of trips is produced by this user, 231, all by car. Furthermore, the destinations varied greatly, 11 destination places in 2 towns and in the city of Seville.

In every dataset the sampling interval fixed in the GPS devices is 5 seconds except user 3 whose sampling interval is 1 second.

The first step in our approach is the filtering of each trip (see Section 2). The results of this procedure for each user are presented in Table 1. It can be

Table 1
Results of the filtered trips and the support points obtained in the experiment.

	Points			Support Points	
Users	Total	Filtered	% of Total	Number	% of Total
User 1	6681	2733	40.91%	150	2.25%
User 2	36161	10517	29.08%	230	0.64%
User 3	100510	13623	13.55%	284	0.28%
User 4	24099	24011	99.63%	545	2.26%
User 5	117820	36216	30.74%	1054	0.89%
User 6	143701	61179	42.57%	1024	0.71%

observed that the filtering of data reduces the number of points, although this reduction is very varied. For example, the percentage of reduction for user 3 is great (86.45%), nevertheless for user 4 it is very small (0.27%) due to the aforementioned particular features of their trips.

It is worth noting the high reduction of points from the log information of each user to those points included in our HMM model: the support points. This is one of the most important advances in our approach since in the most extreme case in our experiment the percentage of reduction is 97.74%. This reduction of points is less if the trips are realized by car than if the trips are by bicycle, or on foot (running or walking). As an example, the major reductions are attained through user 2, 3 and 5 (99.36%, 99.72% and 99.11% respectively) who use other means of transport besides a car. This approach can be useful implemented in a mobile device which carries out the prediction in real time since the number of support points is very small given the nature of the problem studied.

For the prediction of a final destination, given data of only a partial trip, the criteria used to validate the trips and destinations in our experiment was M-fold cross validation on the whole set of training data, where M is chosen according to the size of the dataset, and this procedure is repeated 50 times in order to ensure good statistical behaviour. In every test trip, the support points generated in the explained process are extracted for 25, 50, 75 and 90 percent of the trip travelled (partial trip). These observations allow an HMM model to be generated and after, by using the Viterbi Algorithm [13], the most probable states (destination) for each observation can be obtained. These results are shown in Table 2 and Figure 4.

Table 2
Results of the experiment. M denotes the number of folds in cross validation. The values denote the percentage of correct predictions.

	% Trip traversed			
Users_M	25%	50%	75%	90%
User 1_{10}	36.11%	45.56%	80.56%	88.89%
User 2_{10}	48.81%	58.81%	76.90%	82.62%
User 3_{10}	60.00%	71.36%	83.18%	89.39%
User 4_{10}	64.20%	86.42%	91.85%	94.57%
User 5_5	43.59%	49.04%	67.13%	77.50%
User 6_5	38.57%	55.15%	66.58%	71.81%

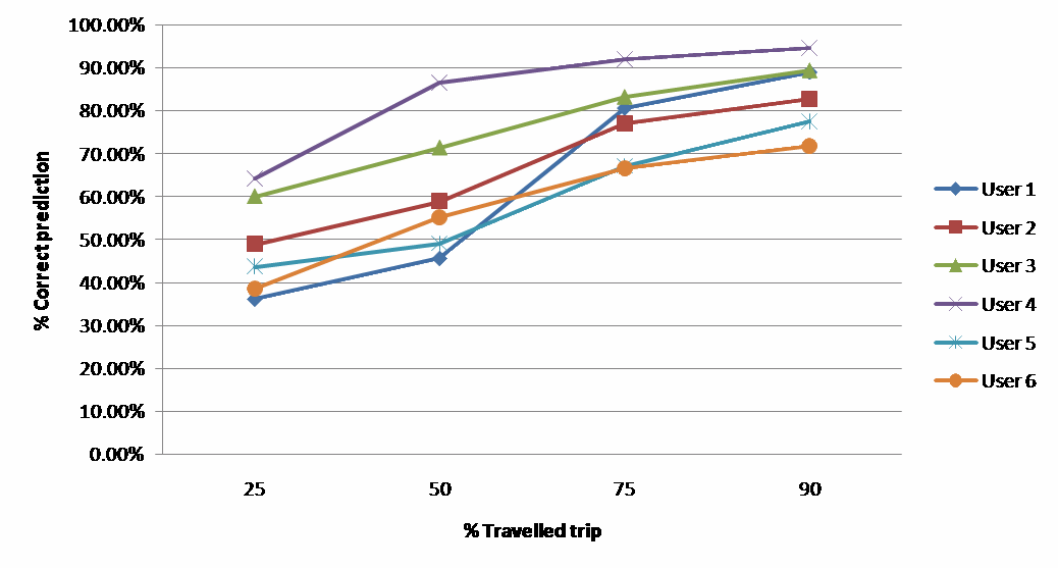

Figure 4. Results of correct predictions from 25% to 90% of the trip travelled .

There are two possible outcomes when a correct prediction cannot be made: Impossible or erroneous predictions. The former occurs when the trip traverses none of the generated support points in the training set. If the zone where the prediction is carried out is a city, then there are very few trips with no observations, but if a wide zone is under study or the pattern of movement uses no trips that cross others, then support points cannot be generated. This situation is solved when the number of trips is greater and therefore more information about the final destination of each user is available.

It can be seen that as the percentage of the trip travelled increases, then the percentage of correct predictions improves. This is a logical result since the number of observations and crossroads is higher and the probabilities of impossible destination predictions decrease. It should be pointed out that,

Although the results are not near 100% of correct predictions, the last support points obtained in 90% of travelled trips are at a point along the route which is significantly less than 90% of the whole trip, in some cases several kilometers, and this fact is due to the method in which the support points have been selected.

With respect to the results, it can be observed in Figure 4 that these vary greatly for each user, and is explained by the different patterns of movements, the zone where they move in and the number of frequently visited places. For example, although there are many more trips by user 5 and 6, they visit places close to each other which implies that the number of correct predictions decreases. It can be seen that user 4 is the most predictable because the same routes are used and the number of final destinations is small. Another curious result is provided by user 1 since it can be seen that the percentage of correct predictions increases from 45.56% (when 50% of the trip is travelled), to 80.56% (when the 75% of the trip is travelled). The explanation of this situation is that the first support points are closer to each other and therefore the prediction is very difficult, but the following points clearly lead to discriminate the final destination.

5 Conclusions

A new approach has been presented to predict destinations given only data of a partial trip by using Hidden Markov Models. Although there are some previous studies related to this approach, they all use a street-map of the city under study and therefore any comparison of results cannot be carried out. It is important to point out that our approach generates its own local street-map and this improves predictions for each user.

In our approach there is a drastic reduction of significant points (ranging from 97.74% to 99.72%) from the first data log information obtained from a user to the resulting set of support points. Hence, the local street-map generated by this approach is simpler than a street-map of the city. Furthermore, the street-map generated is unique for each user.

Due to the reduced number of support points, our "light" HMM can be implemented on mobile devices which include the GPS capability, and can therefore predict destinations accurately in real time, anywhere and without the need of any street-map database. In our experiment, the percentage of correct predictions carried out by the HMM can be considered highly competitive with any other approach.

Acknowledgments

This research is partially supported by the MEC I+D project FAMENET-InCare (TSI2006-13390-C02-02) and the Andalusian Excellence I+D project CUBICO (TIC2141).

References

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

[2] L. E. Baum, J. Eagon, An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology, Bulletin of the American Mathematical Society 73(3) (1967) 360–363.

[3] L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, The Annals of Mathematical Statistics 37 (6) (1966) 1554–1563.

[4] Y. Deguchi, Hev charge/discharge control system based on navigation information, in: International Congress and Exposition On Transportation Electronics, 2004.

[5] J. Froehlich, J. Krumm, Route prediction from trip observations, in: Society of Automotive Engineers (SAE) World Congress, 2008.

[6] J. Krumm, A Markov model for driver turn prediction, in: Society of Automotive Engineers (SAE) World Congress, 2008.

[7] J. Krumm, E. Horvitz, Predestination: Inferring destinations from partial trajectories., in: Ubicomp, 2006.

[8] L. Liao, D. Fox, H. Kautz., Learning and inferring transportation routines, in: 19th National Conference on Artificial Intelligence (AAAI), 2004.

[9] N. Marmasse, C. Schmandt, A user-centered location model, Personal and Ubiquitous Computing (2002) 318–321.

[10] D. Patterson, L. Liao, D. Fox, H. Kautz, Inferring high level behavior from low level sensors., in: Fifth Annual Conference on Ubiquitous Computing (UBICOMP), 2003.

[11] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, in: proceedings of the IEEE, vol 77, no. 2, 1989.

[12] R. Simmons, B. Browning, Y. Zhang, V. Sadekar, Learning to predict driver route and destination intent, in: IEEE Intelligent Transportation Systems Conference, 2006.

[13] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, Information Theory, IEEE Transactions on 13 no 2 (1967) 260–269.