Taxi-Aware Map: Identifying and predicting vacant taxis in the city

Santi Phithakkitnukoon[1], Marco Veloso[2,3], Carlos Bento[2], Assaf Biderman[1], and Carlo Ratti[1]

[1] SENSE*able* City Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Departamento de Engenharia Informática, Universidade de Coimbra, Portugal
[3] Escola Superior de Tecnologia e Gestão de Oliveira do Hospital, Instituto Politécnico de Coimbra, Portugal
santi@mit.edu, mveloso@dei.uc.pt, bento@dei.uc.pt, abider@mit.edu, ratti@mit.edu

Abstract. Knowing where vacant taxis are and will be at a given time and location helps the users in daily planning and scheduling, as well as the taxi service providers in dispatching. In this paper, we present a predictive model for the number of vacant taxis in a given area based on time of the day, day of the week, and weather condition. The history is used to build the prior probability distributions for our inference engine, which is based on the naïve Bayesian classifier with developed error-based learning algorithm and method for detecting adequacy of historical data using mutual information. Based on 150 taxis in Lisbon, Portugal, we are able to predict for each hour with the overall error rate of 0.8 taxis per 1x1 km^2 area.

1 Introduction

We envision a map-based service platform that allows access to real-time information about the state of taxi transportation as well as predictions regarding its future state. A pilot service that turns available taxi-GPS data into useful contextual information that will be provided to citizens for making taxi transportation more efficient and pleasant to use and to policy-makers as a decision-support tool. As an initial step towards building such a platform, we present in this paper a framework for predicting the number of vacant taxis in a given area. The knowledge of the current state of the taxis (number of vacant taxis) in different areas in the city as well as the future state provides the information for a better scheduling. For example, a tourist who arrives at an airport in a transit city and wants to make a trip inside the city with limited time will benefit from the service by using it to plan out a series of taxi rides around the city. A taxi service provider can also use the platform to monitor their taxis and optimize dispatch scheduling. The envisaged platform paves the way for *Ambient Intelligence* (AmI)'s smart city concept that refers to physical environments in which information and communication technologies and sensor systems disappear as they become embedded into physical objects and the surroundings in which we live, travel, and work [1].

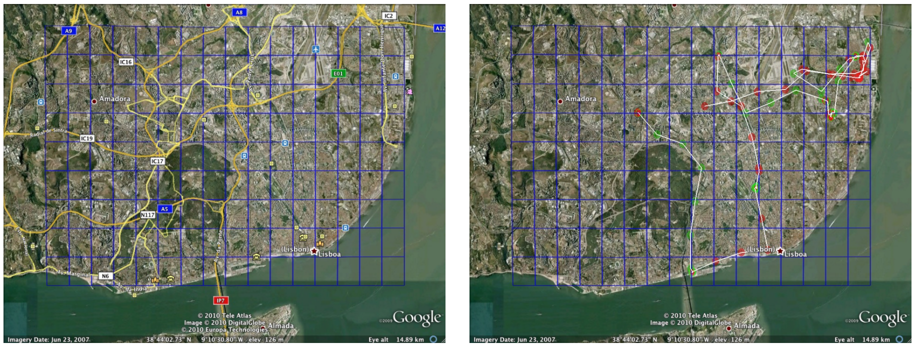
2 Related Work

In recent years, a massive increase in the volume of records of when and where people are has been produced with large deployment of pervasive technologies in the cities. These digital footprints of individual mobility pattern have motivated an increasing number of research in human mobility such as city dynamics [2], predictability of human mobility [3], event-driven traveling pattern [4], and activity-based mobility pattern [5].

GPS-enabled vehicle data such as taxi traces have been collected and analyzed. Chang et al. [6] describe a model that predicts taxi demand distribution based on time, weather condition, and location. Results are shown with different clustering techniques and based on five taxi drivers over two months in service. Yamamoto et al. [7] propose an adaptive routing algorithm using fuzzy clustering to improve taxi dispatching by which vacant taxi drivers are adaptively assigned to pathways with many potential customers expected. Ziebart et al. [8] describe a framework that probabilistically models a distribution over all behaviors (i.e., sequences of actions) using the principle of maximum entropy within the framework of inverse reinforcement learning. With 25 taxis, they show that their model outperform the existing models in turn, route, and destination prediction. Liu et al. [9] analyze 3,000 taxi drivers' operation behavior and categorize them into top and ordinary drivers based on the income. The result reveals the top driver's intelligence through path choice and location.

3 Data Preparation

In this research, we use anonymous data of taxi-GPS enabled traces collected during the period from Sept. 1 thru Dec. 15, 2009 by Geotaxi[10]. This includes 2.6 million anonymous locations of 150 taxis in Lisbon, Portugal. Lisbon is the capital of Portugal with its urban area expanding around the downtown with greatest population density, touristic, historic and commercial areas, and the center for public transportation services. Residential area, airport, and industrial facilities are located in the citys periphery. The data sampling rate varies according to the trip – distance driven, time elapsed, or state changed (e.g. occupancy). Each of the 150 taxi traces carries the information about the taxi's location, service status (occupied, vacant), and the corresponding time and date. If $S = \{s_1, s_2, ...\}$ represents a trace of a taxi, then each instance sample k contains location, service status, and timestamp, i.e. s_k = (latitude$_k$, longitude$_k$, service$_k$, unix timestamp$_k$). For our analysis, we model the map of Lisbon with one-kilometer square grid cells. For each cell, the ID is assigned in raster scanning fashion where it begins with ID #1 at the bottom left corner and increases horizontally left to right. Once it reaches the most right cell, then the counting continues on the next line (upward). The total number of cells in this study is 144 with 16 horizontal cells and 9 vertical cells, as shown in Fig 1(a). To give the readers a sense about our data, Fig. 1(b) shows a sample taxi trace of 10 hours in service, where each dot represents the recorded location with red and green color denoting the service status of being occupied and vacant, respectively.



(a) Area of study

(b) A sample taxi trace

Fig. 1. Map of Lisbon is modeled by 1x1 km^2 grids and a sample of a taxi trace during 10 hours in service. The red dot represents reported location when the taxi is occupied while the green one shows when the taxi is vacant.

4 Predictive Model

The objective is to predict the number of vacant taxis in a grid cell for a given time. To do so, in a probabilistic approach, we need an *inference engine* that estimates the probability of some numbers of vacant taxis within a cell given some observables drawn from historical data. The prediction can then be derived according to the maximum probability criterion. Clearly, if the predictor performed perfectly, one would expect no error. In reality, that is not the case. To improve the performance of the predictor, the predictor thus needs to learn from the errors. Hence *error-based learning* is an essential part of the predictor. One of the most critical elements in context-aware computing and AmI system design is the large data processing as the system must deal with the increasing amount of sensory data to build a priori knowledge for inferring the context of the users. Similarly, in our case, we need a method to detect the *adequacy of historical data*, which will reduce the amount of data in repository and computational cost while retaining the amount of information contained in the original data. In this section, we will describe our approach in building the predictive model that consists of inference engine, error-based learning, and adequacy of data detection. The system overview is shown in Fig. 2.

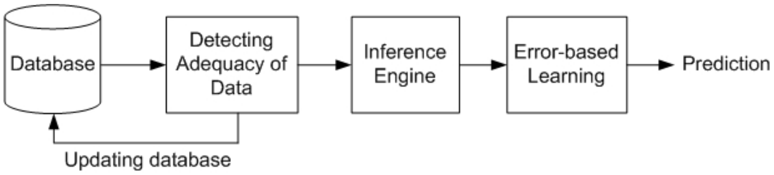


Fig. 2. System overview.

4.1 Inference Engine

Taxi drivers prefer the areas with the most potential customers as well as the areas with potential long-trip customers e.g. airport. A taxi, once occupied by a customer(s) will move to the customer's destination via some path (not necessarily the shortest path). Once a customer ride is completed, the taxi becomes vacant and cruises either in the same area or goes to other areas to seek its next customers. Typically, each taxi attempts to minimize the search time for the next customers. It is intuitive that a taxi driver wants to make the most money in the least amount of time, and hopes to pick up many passengers whose destinations are in places where there are customers to perpetuate a chaining of constant business. As random as it seems, taxis can be predictable to some extent. Their mobility patterns are driven by customers and their individual driving strategies (to maximize revenue). Where and when they become vacant is of our interest in this study. Thus, by considering each grid cell individually, a priori knowledge about the number of vacant taxis of given time can be derived from historical data.

As people normally travel according to the business hours around the city, taxis thus move accordingly to meet the demand e.g. more vacant taxis in residential areas in the morning and in the commercial areas in the afternoon. *Time of the day* therefore becomes a reasonable indicator for estimating vacant taxis in an given area. Besides time of the day, *day of the week* seems to carry some clue about the vacant taxis as well. For example, there tends to be more taxis in business areas during the weekdays than in the weekends. *Weather condition* also plays an important role in traveling decision making. A nice sunny day may attract more people to go out and travel compared to a rainy day. The number of vacant taxis thereby varies with the characteristic of the given area captured by a series of observations based on these factors.

Our inference engine is constructed with a *naïve Bayesian classifier*, which is a probabilistic classifier based on Bayes theorem with independence assumptions [11]. In our case, we want to compute the likelihood of each possible number of vacant taxis (Y) given time (hour) of the day (T), day of the week (D), and weather condition (W). The probability of the number of vacant taxis is y_i can be computed as

$$P(Y = y_i|T, D, W) = \frac{P(Y = y_i)P(T, D, W|Y = y_i)}{P(T, D, W)}, \tag{1}$$

where $T = \{1, 2, 3, \ldots, 24\}$, $D = \{\text{Monday, Tuesday, \ldots, Sunday}\}$, and $W = \{\text{Sunny, Cloudy, Rainy}\}$. The prediction is then made using MAP method [11], which selects the number of vacant taxis that (y_{MAP}) that maximizes *a posteriori* as follows:

$$\begin{aligned} y_{MAP} &= \arg\max_{y_i \in Y} P(Y = y_i|T, D, W) \\ &= \arg\max_{y_i \in Y} P(Y = y_i)P(T, D, W|Y = y_i) \\ &= \arg\max_{y_i \in Y} P(Y = y_i) \prod_i P(T|Y = y_i)P(D|Y = y_i)P(W|Y = y_i). \end{aligned} \tag{2}$$

As an example, Fig. 3 shows distributions of vacant taxi volume given different time of the day, day of the week, and weather condition. This sample is based on the history of grid #28. The information of weather condition is obtained from Weather Underground [12].

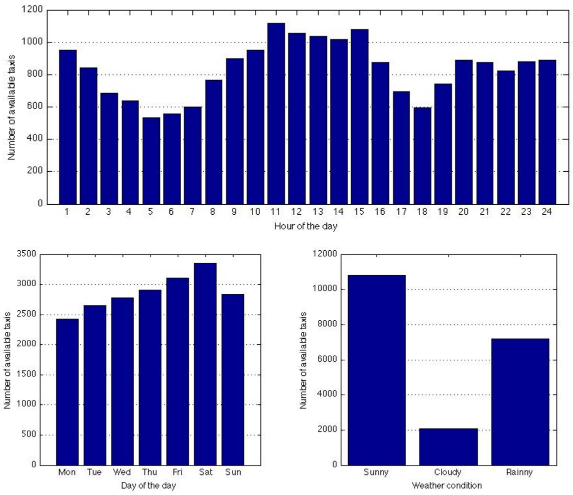


Fig. 3. A sample distribution of vacant taxis given time of the day, day of the week, and weather condition. This sample is drawn from grid #28.

4.2 Error-based Learning

With the inevitable uncertainty that any predictive model must deal with, the actual outcome may not occur as predicted. Hence error occurs in the system. Instead of letting the error occurs while using it as part of evaluation for the model, we utilize the error to improve the performance of our predictor. Especially the errors that take place most recently indicates the possibly change in pattern. An efficient system must be able to automatically detect and adapt into the change. So the key is to learn from the "recent" errors as the recent data reflects the current trend. Our approach is to apply a weight function that emphasizes on the recent errors from which the prediction is then adjusted accordingly. One possible weight function is the *uniform* function as given by Eq. (3).

$$w_u(t) = \begin{cases} \frac{1}{\beta} & t_m - \beta \leq t \leq t_m \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where t_m is the most recent time and β is the bandwidth (indicating how much of history being taken into account). The weight is distributed equally over the most recent β errors. Another possibility is to assign more weight for more recent errors within the bandwidth. This can be a *linear* function such as the one given by Eq. (4).

$$w_l(t) = \begin{cases} \frac{1}{\beta}(t - t_m + \beta) & t_m - \beta + 1 \leq t \leq t_m \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The prediction will therefore be adjusted according to the weight function as:

$$\hat{y}(t) := \hat{y}(t) + \sum_{i=t-\beta}^{t-1} w(i)e(i), \tag{5}$$

where $\hat{y}(t)$ is the prediction made at time t, $w(i)$ is the weight function, and $e(i) = y(i) - \hat{y}(i)$ is the error measured by the difference between the actual $(y(i))$ and predicted value.

4.3 Adequacy of Data

With the vision of the AmI, which refers to the environment equipped with surrounding computing devices that are sensitive and responsive to the presence and context of people, designers of such environment must deal with the processing of a large human behavioral data. These data can be collected from multiple sensors and processed in some sophisticated way to extract some knowledge about the users. The larger data becomes, the higher capacity is required for storage as well as computation. If the goal is to capture the core structure of some pattern in the data, then the entire data might not be necessary needed. The interesting question is then *how much of historical data is actually adequate to characterize behavioral pattern of interest?*

In our case, the historical data is used to construct a priori distributions for the predictor. Thus, the amount of historical data that allows us to capture the same distributions as we were to use the entire data is "adequate". The behavioral pattern of interest in our case is therefore the distribution of vacant taxis over the grid cells.

To detect the adequacy of historical data, we apply information theory's mutual information [13], which is a measure of the amount of information that one random variable contains about another random variable. Let X denote a random variable representing entire data and Z be a random variable representing some amount of the most recent data in X. The mutual information $I(X;Z)$ is defined as the reduction in the uncertainty of X due to the knowledge of Z, as follows:

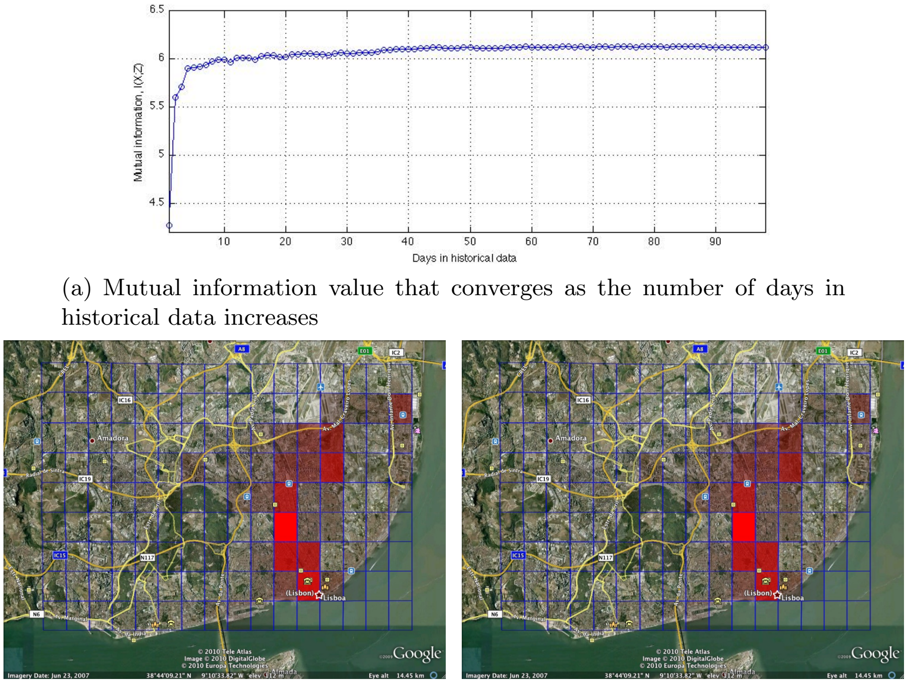
$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &= H(X) + H(Z) - H(X,Z), \end{aligned} \tag{6}$$

where $H(X)$ and $H(X,Z)$ are information entropy (uncertainty) of X and joint entropy given by Eq. (7) and (8), respectively.

$$H(X) = -\sum_x p(x) log_2 p(x) \tag{7}$$

$$H(X,Z) = -\sum_{x,z} p(x,z) log_2 p(x|z) = -\sum_{x,z} p(x,z) log_2 \frac{p(x,z)}{p(z)} \qquad (8)$$

The idea is to find the amount of data carried by Z that can describe most majority of information contained in X. In our case, $p(x_i) = n_i / \sum_i n_i$ where n_i is the amount of vacant taxis in cell i. Based on our experiment, in turns out that a certain amount of historical data carried by Z is adequate to characterize X. As an example, Fig. 4(a) shows the value of $I(X;Z)$ as Z continues to carry more historical data (in days). The value of $I(X;Z)$ converges around 40 days of historical data. Since $I(X;X)$ is equal to $H(X)$ (self-information), the result implies that $I(X;Z) \approx I(X;X) = H(X)$ when at least the last 40 days of data has been taken into account. In the other words, the last 40 days of data is adequate to describe the entire data. Figure 4(b) and (c) show similar distributions of vacancy derived from the entire data and the last 40 days, respectively. This example is one of the taxi traces in our dataset.



(a) Mutual information value that converges as the number of days in historical data increases

(b) Distribution of entire data

(c) Distribution of retained data before convergence (40 days)

Fig. 4. A sample from Taxi ID #9 of convergence of mutual information values and distributions of entire data as well as retained data of recent 40 days. The color shade represents the value of the distribution over different grids.

5 Experimental Results

To test the performance of our predictive model, the first 30 days of data is used for training the model while the rest of the data (77 days) is used as a testing set. A prediction of the number of vacant taxis is made for the next 24 hours (one for each hour slot – 24 predictions are made for each testing day) for each grid cell. The errors are then computed and used to adjust the later predictions (as described in Sect. 4.2). The testing data will sequentially become training data after each prediction has been made. Adequacy of historical data is detected each day of testing (as described in Sect. 4.3) and the training set is then updated accordingly. To detect the convergence time, we use a method described by Phithakkitnukoon and Dantu [14] with threshold of 0.02.

During the experiment, the absolute error is computed as the absolute difference between the actual and predicted amount of vacant taxis, i.e. $|\hat{y} - y|$. The cumulative error is computed over all grid cells and shown in Fig. 5. To show the improvement of the predictor achieved by the error-based learning and detection of adequacy of data, in Fig. 5, cumulative errors of the model with (full model) and without these features (baseline model) are shown, where clear improvement of about 10% in error can be observed. The uniform and linear weight functions are used (with $\beta = 5$) where both functions yield similar performance. The overall error per grid cell is shown in Fig. 6 where cells in blue, yellow, and red have error range of 0–1, 1–2, and 2–3 taxis, respectively. The higher error cells seem to be clustered in commercial and touristic areas while lower error cells are spreading gradually. Overall, the errors are in the range from zero to three taxis, which is considerably promising. In the other words, the reliability of each prediction is $\pm c$ where $c = 1$, 2, and 3 if the cell is blue, yellow, and red, accordingly.

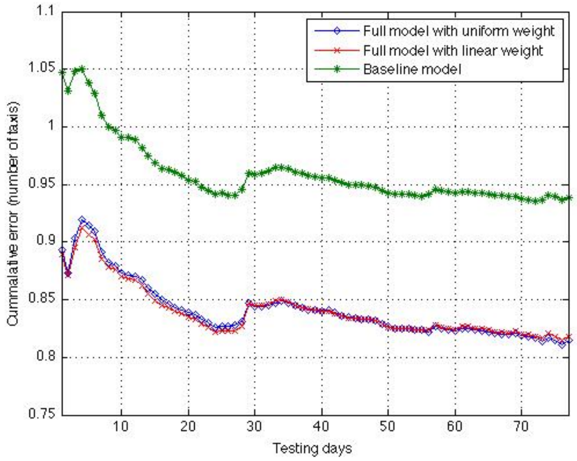


Fig. 5. Cumulative error over days in testing period. The full model with the error adjustment and adequacy of data outperforms the baseline model.

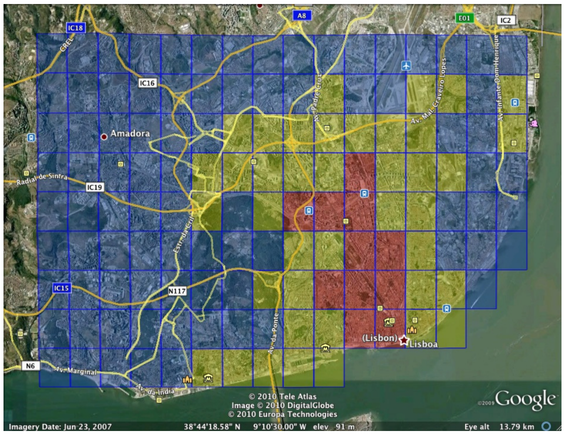

IC18
A8
E01
IC2
IC16
Amadora
Radial de Sintra
IC19
IC15
N117
N6
(Lisbon)
Lisboa
© 2010 Tele Atlas
Image © 2010 DigitalGlobe
© 2010 Europa Technologies
Google
Imagery Date: Jun 23, 2007
38°44'18.58" N 9°10'30.00" W elev 91 m
Eye alt 13.79 km

Fig. 6. Error rate in different grid cell. Error ranges are 0–1, 1–2, and 2–3 taxis, which are represented with blue, yellow, and red respectively.

6 Conclusions

With the vision of smart cities, which is an application of concept and development in AmI, here we take the first and essential steps toward building a map-based service platform that allows access to real-time information about the state of taxi transportation as well as predictions regarding its future state. This will be useful for taxi service providers, transportation management, as well as taxi users. In this paper, we develop a framework for predicting the number of vacant taxis in an area based on time of the day, day of the week, and weather condition. Our predictive model is based on Bayesian classifier with a sequential error-based learning algorithm and a mechanism for detecting the adequacy of historical data using information theory's mutual information. Based on 150 taxis in Lisbon, Portugal, our model is able to predict (per hour) accurately with the overall error of slightly over 0.8 taxis per 1x1 km^2 area. As our future direction, we will continue to find ways to improve our framework such as by using different weight functions (e.g. exponential, sigmoid), weight bandwidth, and exploiting multi-source data fusion (combing data from different sources e.g. mobile phone, bus, fleet). A larger dataset will also be obtained (from Geotaxi) for our future development.

Nonetheless there is a number of limitations of this study that we are aware of as following:

- The sample size of 150 taxis may not be the true representative of the whole taxi population. It would be interesting to see how our framework performs with a larger sample size.
- The grid size is relatively large and may not be realistic for individual users who are particularly interested in the prediction within a smaller area. In this study, the grid size is designed to compensate the taxi sample size that

we have. With a larger sample size, smaller grid size can be implemented with the same developed framework. Nevertheless, the current grid size is still useful for taxi service management and urban/transportation planning.

- The weather condition is assumed to be the same throughout the day. For example, if the record shows that September 1st, 2009 was a sunny day, then the weather condition is assigned to be sunny for entire 24 hours, which may not be true. Nonetheless, we do not anticipate a large change in the overall result if we had accounted for the change of weather condition during day.

Acknowledgments. The authors gratefully acknowledge support by Volkswagen of America Electronic Research Lab, the AT&T Foundation, the National Science Foundation, MIT Portugal, and all the SENSEable City Laboratory Consortium members.

References

1. Steventon, A., Wright, S.: Intelligent Spaces: The Application of Pervasive ICT (Computer Communications and Networks). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005)
2. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing **6**(3) (2007) 30–38
3. Song, C., Qu, Z., Blumm, N., Barabsi, A.L.: Limits of predictability in human mobility. Science **327**(5968) (2010) 1018–1021
4. Calabrese, F., Pereira, F.C., Lorenzo, G.D., Liu, L.: The geography of taste: analyzing cell-phone mobility and social events. In: Proceedings of IEEE Inter. Conf. on Pervasive Computing (PerComp). (2010)
5. Phithakkitnukoon, S., Horanont, T., Lorenzo, G.D., Shibasaki, R., Ratti, C.: Activity-aware map: Identifying human daily activity pattern using mobile phone data. In: Inter. Conf. on Pattern Recognition (ICPR 2010), Workshop on Human Behavior Understanding (HBU), Springer (2010) 14–25
6. Chang, H., Tai, Y., Hsu, J.Y.: Context-aware taxi demand hotspots prediction. Int. J. Bus. Intell. Data Min. **5**(1) (2010) 3–18
7. Yamamoto, K., Uesugi, K., Watanabe, T.: Adaptive routing of multiple taxis by mutual exchange of pathways. Int. J. Knowl. Eng. Soft Data Paradigm. **2**(1) (2010) 57–69
8. Ziebart, B.D., Maas, A.L., Dey, A.K., Bagnell, J.A.: Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In: UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing, New York, NY, USA, ACM (2008) 322–331
9. Liu, L., Andris, C., Bidderman, A., Ratti, C.: Revealing taxi drivers mobility intelligence through his trace. Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches (2010) 105–120
10. Geotaxi. http://www.geotaxi.com
11. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
12. WeatherUnderground. http://www.wunderground.com/
13. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York, NY, USA (1991)
14. Phithakkitnukoon, S., Dantu, R.: Adequacy of data for characterizing caller behavior. In: Proceedings of KDD Inter. Workshop on Social Network Mining and Analysis (SNAKDD 2008). (2008)