Habitualisation: localisation without location data

Rory McGrath
National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland

Cathal Coffey
National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland

Alexei Pozdnoukhov
National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland
Alexei.Pozdnoukhov@nuim.ie

ABSTRACT

This paper looks at identifying the locations of users from the Nokia MDC dataset throughout the day without taking into consideration location based data. By looking at a users habits and idiosyncrasies we determined the likelihood of a users location within known stay regions which we call habitats. The features used to determine location were extracted from a users interaction with the smart phone. None of the features contained a users locations or a users proximity to objects with known locations. Using a set of structured output support vector learning techniques we found that a users location with respect to the areas of typical activities is well predictable solely from daily routines and a smart phone usage habits.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition—*Neural nets*; H.2.8 [**Database Management**]: Database Applications—*data mining, mining methods and algorithms, interactive data exploration and discovery*

General Terms

ALGORITHMS

Keywords

machine learning, kernel methods, smart cities, pervasive computing

1. INTRODUCTION

Location prediction remains the keystone of traditional location-based services operating on mobile devices. However a blind belief into positioning technologies which solely answer the "where" question may discredit the convenience of the services pushed to the user if location is taken out of context of previous actions or likely future intentions. Location can not be separated from activities: a lunch with colleagues on a weekday is not the same as a family dinner at the same restaurant on a weekend.

The data on irregularly locations of individuals either via cell phone usage logs [11], "check-ins" in location-based social network services [8] or geo-tagged Twitter texts [10, 9], as well as databases of high precision GPS traces are becoming available [13]. Travel routines observed in these data are mainly daily commutes which dominate mobility pattern [11]. Predictive modelling of human movement beyond regular commutes is a challenging task. Empirical observations [1, 10, 3] suggest the importance of social influence on the formation of atypical patterns of mobility. People tend to follow recommendations of their friends in planning travel, or joining them on a trip to explore new areas

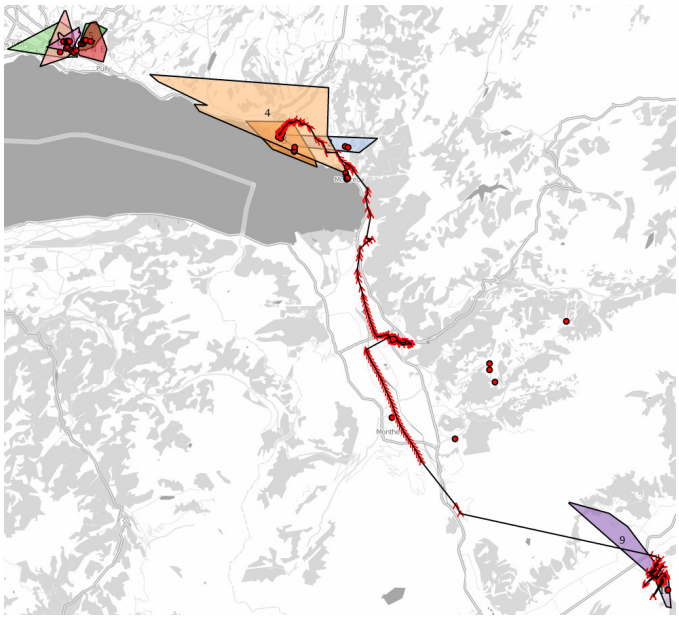

Figure 1: Example of a users trajectory, stay points and estimated GSM cell coverage.

Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing Newcastle, UK, June 2012

and visit particular places for recreation, leisure or tourism. Some studies suggest the presence of the so-called "habitats" [2, 5], formation of which is related to spatial choice processes in human decision making and contextual activities.

A distinct feature of the Nokia MDC dataset is that it provides both GPS locations and logs of user interaction with the phone thus providing contextual features to enhance predictive models of users whereabouts.

1.1 Contributions of the paper

In this paper we considered location prediction as a structured output learning problem in the context of activities captured by a phone usage log. We have applied an SVM-HMM model [12, 4] to train a model for user locations from both temporal patterns in the sequence of visited locations and contextual features of phone use. We have modelled spatial movement as a walk process over a graph of important locations. Following the idea of habitats [5], we built a hierarchy of locations based on the spatial resolution and availability of locational footprints of movement patterns of a user. We simplified the graph structure to 'habitats' due to two main reasons. First, we built a model that unifies locational and phone usage sequences into sequences of locational/behavioral habitats. Second, we increased predictability performance by decreasing the number nodes and increasing the number of transitions, however, keeping it well beyond a trivial bi-modal structure of home-to-work commute (Section 3.1). Input features were generated from a users' interaction with the phone and included a variety of modalities described in Section 3.3. We applied a range of support vector sequence learning methods and prediction results on a held-out testing dataset are reported in Section 4 and discussed in Section 5.

2. TECHNIQUES

There is a rich framework of methods developed in the area of human activity recognition within computer vision and smart environment application areas. The simplest category of activity recognition models ignores temporal dependencies assuming subsequent inputs to be independent, and encompass a wide selection of classification methods (support vector machines (SVM), naive Bayes classifier, nearest neighbors) which can be used baselines. The second category originates at Markov chains background. These are hidden Markov models (HMM), conditional random fields with latent Markov chains, and, particularly, SVM-HMMs [12] that consider statistical dependencies over adjacent feature vectors and show good performance on pre-segmented data with high-dimensional input vectors.

3. DATA PROCESSING

3.1 Locations data

Before we could predict locations for the users we first had to find the areas that they most frequently visited. To find interesting geographical points for our users we estimated a set of stay points from each users GPS trajectories using an algorithm proposed by [6]. This algorithm has two parameters: a distance threshold in meters and a time threshold in seconds. These parameters are used to estimate a collection of locations, with a maximum area of the distance threshold, that a user has stayed at for a minimum of the time threshold. Using this algorithm all spatial regions smaller than a radius of 200 meters where a user spent at least 30 minutes within were found.

After calculating these stay points they were clustered into stay regions using grid based clustering. This was carried out to create stay regions which contained all stay points that had the same semantic meaning but due to GPS error have slightly different centroids. This clustering technique was also used for the same dataset [7].

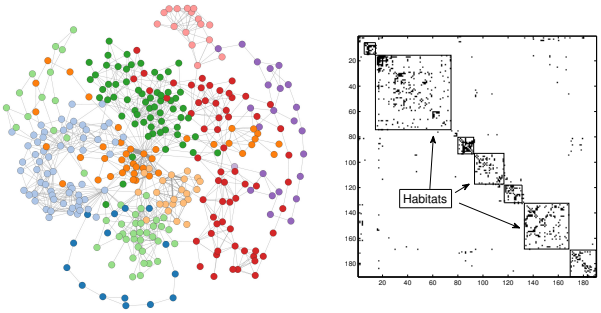

Figure 2: Undirected transitions graph between stay points colored by habitat category (left), and the block structure (the 'habitats') in the transitions matrix identified via community detection (right).

The stay regions were then visualised as an undirected network. From this we observed that there was a limited number of transitions between each stay region. The information relating to GSM towers was recorded by the phone every few minutes and provided the id of the currently connected tower. The data set did not provide cell tower locations or estimated regions of broadcast. To estimate these all of a user's GSM and GPS data were combined. A GPS was defined as belonging to a GSM cell if a GPS point had high accuracy (accurate to 50 meters) and the user had a strong cell phone signal (strength = 7) and the time difference between GPS and GSM clocks was less than 5 minutes. A convex hull was then computed for these points to estimate the cell towers broadcast area.

3.2 Habitats identification

The subsets of regions for a users locations of interest were identified by applying community detection as described in [5]. The number of transitions between each tower was calculated by a total count of transitions identified from GPS tracks and GSM towers visited in the dataset. A particular problem was the noise added by phones continuously switching between towers. We used a threshold based on the number of active days to eliminate the effect of "fictitious transitions. This threshold was chosen as it was assumed unlikely that a person would transition between two towers that amount of times during those days.

Once this procedure was carried out, the 13 habitat clusters with the most transitions where then chosen. An example of the graph is visualised as an undirected graph (Figure 3). While these nodes had strong connections with nodes in the complete graph they did not necessarily have strong connections between each other. As we are dealing with a classification problem we need at minimum link strength of two which would allow the nodes to be placed in both the training and testing set.

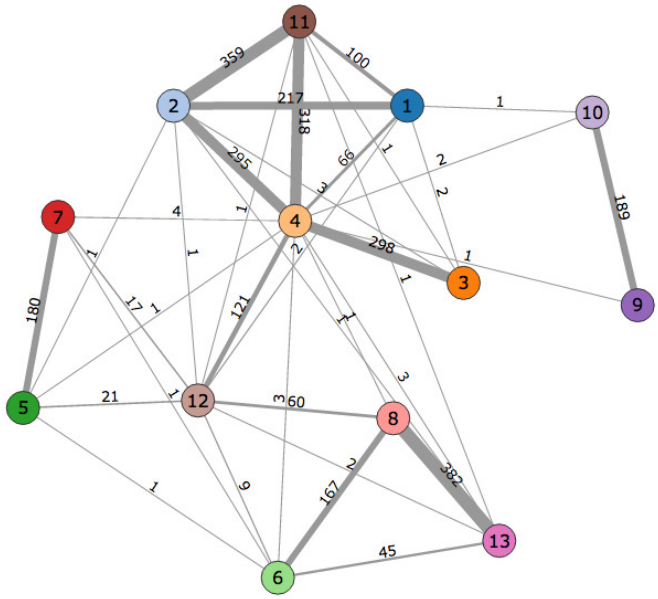

Figure 3: Undirected graph of transitions between habitats identified by a combined use of GPS and GSM towers localisation.

3.3 Behavioral features

The nokia data set provided a large amount of data relating to a users phone usage activities. From this data set features which best describe a users activities were extracted.

When picking features we only looked at properties that the user had complete control over. For this reason we choose to generate features relating to calls and sms made but not for call or messages received. We are interested in predicting locations based on the users habits, while a user may receive a call in most places it is more informative to look at where he makes calls. Does he always go to the same quite place? Does he always make a call when he is at lunch? Does he always text at work?

Binary encoding was used for generating feature values. The feature vector was generated each time a user entered a location and spanned the length of time a user spent there. Fifty-two features were generated to capture the following phone activities:

- **Charging pattern.** Using the charging information we were able to determine whether or not a user was charging their phone and the current state of the charging process. This was described using four features which related to the states 'not charging', 'charging', 'charged but still plugged in' and 'stopped charging for a brief period and resumed charging'.
- **Music.** This binary feature showed whether or not a user was listening to music.
- **Calendar.** This feature was used to show if a user had made an entry into there calendar on the phone.
- **Apps.** Five features were generated to describe if a user was using one of their top five most frequently used applications.
- **SMS.** By looking at the call log four features were generated relating to sms messages sent. These features captured if a user sent a message to one of their top three most frequency contacts or any other contact.
- **Calls**. Four features were also generated to describe all outgoing calls made by the user. Similar to the sms messages this feature captured if a user called one of their three most frequent contacts or if they call anyone else.
- **Contacts.** A feature relating to the users contact information was also generated. This feature showed if a user made a new contact entry.
- **Media.** Looking at the media on the phone a feature was generated to see if any new media was added to the phone.

3.4 Training And Testing Sets

Training and testing sets were generated. These sets were generated such that every node appeared at least once in each. Once one node was present in each set the remaining nodes were split amongst the sets with a ration of 80:20 in favor of training. A ratio of 80:20 of the occurrences of the nodes was also kept between training and testing where possible.

4. EXPERIMENTS

The location of a user was predicted using the SVM, MCSVM and HMSVM techniques.

The accuracy of each model was determined using Cohen's Kappa measure. This measure took into account correctly classified labels as well as correct classifications obtained by chance. Using this score the optimal input parameters to the models were calculated. Using the training data a range of values were used for the input parameter relating to slack vs magnitude of the weight-vector and the corresponding kappa score was recorded. It was observed that a value of 4000 worked best of HMSVM and 16000 for MCSVM in terms of cross-validation performance.

The models were then trained using these tuned parameter and tested using the testing set. The confusion matrices for each model was then generated.

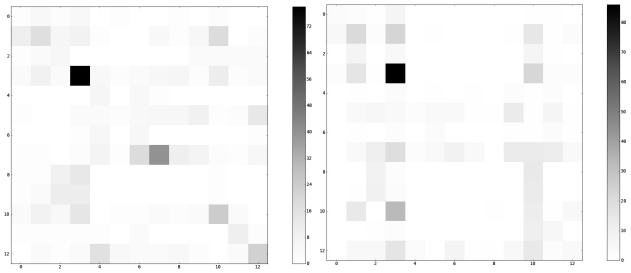

Figure 4: Confusion Matrices. Left: HMSVM Right: MCSVM. The rows show the actual label and the column shows the predicted label

Using the kappa score the HMSVM model performed 1.23 times better than MCSVM. This is also evident in the confusion matrix which shows HMSVM classifies 15% more accurately than MCSVM. The f1 scores were calculated for both methods and the results are presented in table 1.

	f1		Prec		Recall	
Label	m1	m2	m1	m2	m1	m2
1	0.05	0.24	0.04	0.30	0.07	0.20
2	0.28	0.28	0.34	0.27	0.24	0.29
3	0.16	0.00	0.12	0.00	0.24	0.00
4	0.58	0.51	0.57	0.42	0.60	0.66
5	0.22	0.00	0.14	0.00	0.55	0.00
6	0.07	0.13	0.18	0.31	0.04	0.08
7	0.19	0.00	0.12	0.00	0.42	0.00
8	0.51	0.08	0.66	0.57	0.42	0.04
9	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00
11	0.37	0.13	0.38	0.10	0.36	0.18
12	0.45	0.21	0.42	0.16	0.50	0.30
13	0.38	0.10	0.43	0.20	0.34	0.06

Table 1: Table of f1, precision and recall scores obtained. m1 represent the HMSVM model and m2 represent the MCSVM model

From Table 1 we can see that prediction worked well for labels 4, 8, and 13. Label 4 corresponds to the persons home location and label 8 and 13 corresponded to work locations. These locations are frequently visited which explains the predicted.

The labels Labels 9 and 10 are related to a weekend location. The number of transitions between home and weekend or work and weekend are limited as evident in figure 3. This explains the poor performance in prediction.

5. DISCUSSION AND CONCLUSIONS

From the results presented we have shown that a persons transitions between spatial habitats is predictable solely from daily routines and there smart phone usage habits. However, our method does not take into account long-range or global characteristics such as the interactions between activities. This drawback can be overcome with a model which is an extension of HMSVM that considers the underlying process to be a semi-Markov chain (SMM) with a variable duration for each state.

The results were also limitation due to the low number of transitions and repeated sequence chains. Additionally prediction results could be improved by adding accelerometer data to the feature set. This data could be used to indicate the type of activity the user was taking part in. This activity inference along with the users behaviour could potentially improve the accuracy of location prediction.

In conclusion these results enhance our understanding of the fundamental laws of human mobility and can be used to improve quality of service of traditional location-based services and empower the growing popularity of location-based social networks with smart context-aware predictive capabilities.

Acknowledgments

Research presented in this paper was funded in part by Science Foundation Ireland Strategic Research Cluster grant 07/SRC/I1168 and 11/RFP.1/CMS/3247 award, and IBM PhD Fellowship program. The authors gratefully thank Aonghus Lawlor and Felix Kling for their support, fruitful discussions and help with software.

6. REFERENCES

[1] K. W. Axhausen. Social networks, mobility biographies, and travel: survey challenges. *Environment and Planning B: Planning and Design*, 35:981–996, 2008.

[2] J. P. Bagrow and Y.-R. Lin. Spatiotemporal features of human mobility. Feb. 2012.

[3] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[4] T. Joachims, T. Finley, C.-N. Yu, T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, Oct. 2009.

[5] A. Lawlor, C. Coffey, R. McGrath, and A. Pozdnoukhov. Stratification structure of urban habitats, June 2012. Pervasive Urban Applications workshop at PERVASIVE'12.

[6] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 34:1–34:10, New York, NY, USA, 2008. ACM.

[7] R. Montoliu and D. Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, pages 12:1–12:10, New York, NY, USA, 2010. ACM.

[8] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *arXiv:1108.5355v4 [physics.soc-ph]*, 2011.

[9] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[10] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 723–732, New York, NY, USA, Feb. 2012. ACM.

[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, Feb. 2010.

[12] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[13] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Y. Ma. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008. ACM.