Nokia Mobile Data Challenge: Predicting Semantic Place and Next Place via Mobile Data*

Ying Zhu Yong Sun Yu Wang
Department of Computer Science, University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, North Carolina, USA
{yzhu17,ysun20,yu.wang}@uncc.edu

ABSTRACT

The appearance of smartphones equipped with various sensors enables pervasive monitoring of mobile users' behaviors and mobility. The Nokia Mobile Data Challenge (MDC) [10] gives us a great opportunity to study the users' mobility models and location profiles from a rich mobile dataset. The realistic data analysis may benefit a wide range of fields from technology innovation to policy making. In this paper, we describe our proposed methods to predict the semantic meaning of the "important places" (Task 1) and the users' next destination (Task 2) based on released MDC data. For Task 1, we explore several features from the sequence of visited places and accelerometer samples, and proposed two types of prediction methods: rule based and machine learning based. For Task 2, we adopt a simple but effective machine learning method to accomplish the prediction assignment via both temporal and spacial features. Our preliminary experimental results over released MDC data (Set A dataset) show that rule based methods produce good predictions for home and work locations in Task1, while machine learning methods are more scalable for different types of prediction tasks. But the prediction accuracy of machine learning approaches heavily depends on the number of instances in the training data.

1. INTRODUCTION

Inferring the meaning of the most significant places that a mobile user visits frequently and predicting the future location of the user are central to understand human mobility and social patterns. Such understanding can inform solutions to challenging problems in a wide range of fields, such as mobile recommendation systems [16, 17], wireless routing [7, 18], urban sensing and planning [2, 13], sociology [3, 4], ecology and epidemiology [5]. There has been numerous studies on significant location and movement prediction using GPS coordinate data [1, 11, 12, 14, 16, 17]. Recent advances of smartphones equipped with various sensors and contact/event logs enable new directions to study mobile users' behaviors and mobility [3, 4, 8, 9] far beyond GPS tracing.

The Nokia Mobile Data Challenge (MDC) dataset [10] holds great potential for providing fine-quality information to predict the semantic places and user's next destination. From the study of MDC dataset and the ground truth data, we find that people's access of certain place may follow some regular patterns. For example, people usually go to work place during the daytime and go home at night. These patterns are helpful to our prediction tasks. In addition, people's behaviors at some specific places also provide useful information for certain predictions. For instance, if a person is doing an outdoor sport (such as hiking), he/she must have certain speed. But if this person is having a dinner in a restaurant, he/she most likely is stable during the dining period. We believe by that exploring these types of features from the MDC dataset, we could accomplish our prediction tasks. Last but not least, many features are *time-dependent* and *user-dependent*. This is especially true for the next place prediction. We consider both the temporal and spacial features in our proposed prediction methods.

In the following sections, we describe our proposed methods for semantic place prediction (Task 1) and next place prediction (Task 2) in detail. To test the performance of our proposed methods, we conduct a few experiments over the released MDC dataset (Set A in [10]). We include some of these preliminary results in each section.

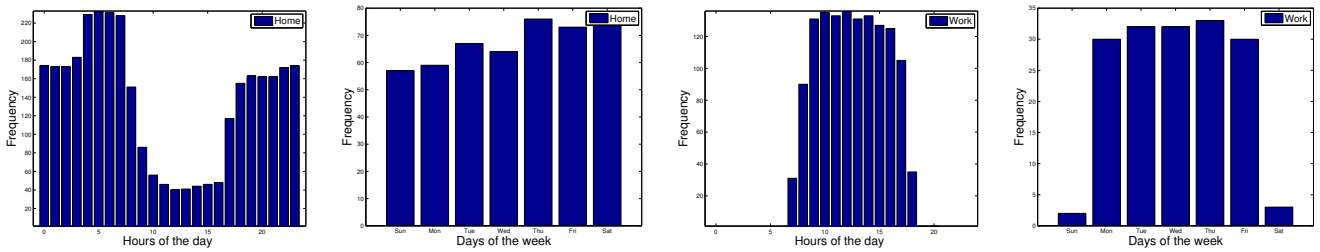
2. TASK 1: PREDICTING SEMANTIC PLACE

In this section, we first introduce the features that we extract from the mobile data for MDC prediction tasks before we describe our proposed prediction methods for Task 1, predicting semantic meaning of places.

2.1 Feature Extraction

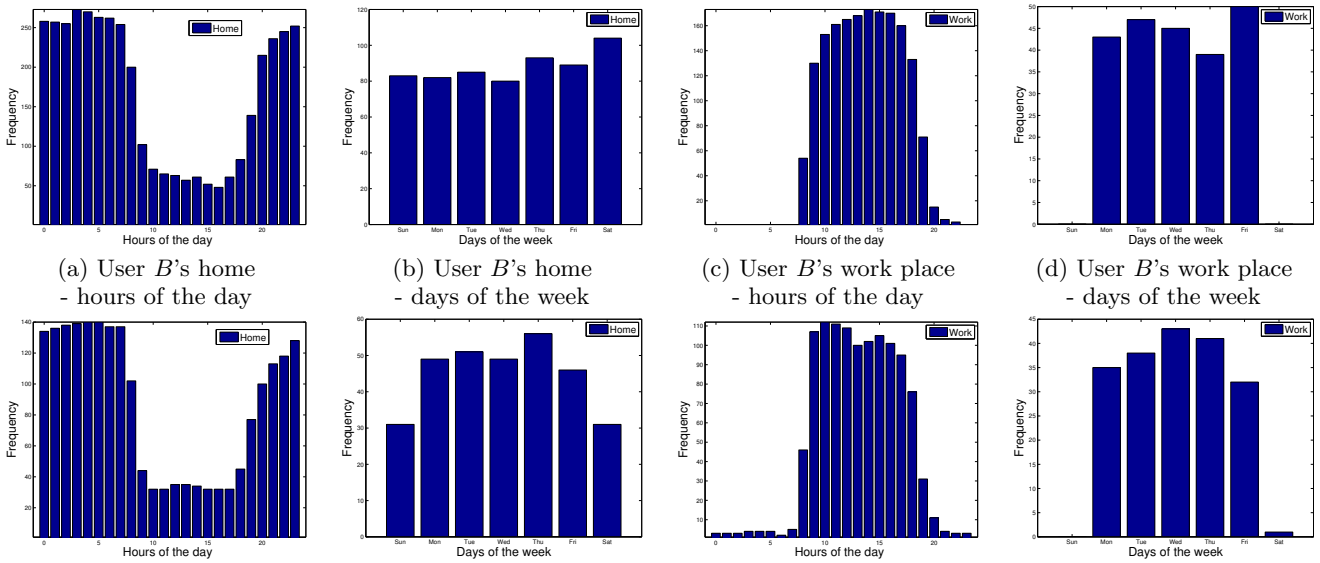
We explore two types of features for Task 1: features from the sequence of visited places ($visit\_sequence\_10min.csv$) and features from scanning data of accelerometer sensors ($accel.csv$).

*This material was prepared for the Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK. The copyright belongs to the authors of this paper. This work was supported in part by the US National Science Foundation (NSF) under Grant No.CNS-0915331 and CNS-1050398.



(a) User A's home - hours of the day (b) User A's home - days of the week (c) User A's work place - hours of the day (d) User A's work place - days of the week

Figure 1: User A's access frequency distribution of his home and work place shows clear patterns: home has higher access frequency at night and during the weekend, while work place has higher access frequency at daytime and during weekdays.



(a) User B's home - hours of the day (b) User B's home - days of the week (c) User B's work place - hours of the day (d) User B's work place - days of the week

(e) User C's home - hours of the day (f) User C's home - days of the week (g) User C's work place - hours of the day (h) User C's work place - days of the week

Figure 2: User B and C's access frequency distribution of their home and work places

People often access particular places with regular patterns. For example, during weekdays, Bob gets up at 7:00am and sends his daughter to kindergarten at 7:30am, and then goes to work around 8:30am. At 5:00pm, he takes off for his office and picks up his daughter, and then stays at home during night. He repeats this routine almost every weekday. Detecting such regular location pattern over time is crucial to distinguish semantic places. For example, from 8:30am to 5:00pm Bob probably is not at home because it is his work time.

Bob's story may be too ideal, however, we do find regular patterns in MDC dataset, especially for home and work places. Figure 1 illustrates the access frequency distribution of a MDC user (let us call him A) for his home and work place during a day and a week respectively. It's obvious that A's access distributions of his home and work places have significant differences. His home has high access frequency at night and equal access frequency over every day, meanwhile his workplace has high access frequency at daytime and during weekdays (Monday to Friday). This implies that features of people's access on a place could be used to predict the semantic meaning of the place.

Even better, different people may share similar access patterns over a particular type of place. Figure 2 illustrates the access frequency distributions of the home and work places of two additional MDC users. They both share the similar regularity with user A. Therefore, for certain types of places, there exists common features and rules to predict them. For example people go to restaurant usually during 12:00am-2:00pm and 5:00pm-8:00pm, and most of people sleep at night. By revealing these observable features, we may successfully predict the semantic meaning of a place.

Via studying ground truth data, we find several time-related observable factors that capture characteristics, which are useful to distinguish the places' semantic meanings. From

the sequence of visited places (*visit_sequence_10min.csv*) of MDC data, we can extract the following features:

- *Number of access days in a month*: the number of days that the user accessed the place within one mouth. Multiple accesses within the same day is only counted for one. This feature mainly helps us to distinguish people's home and work places, from the observation that people almost go to work and go home every day.
- *Total access frequency/duration*: the total times/duration of the place visited by the user during the whole data collection period. This feature implies the importance of such a place.
- *Total access frequency/duration of visits shorter than 2 hours*: the total times/duration of visits of the place shorter than 2 hours. People normally do indoor sports and stay in the restaurant for less than 2 hours.
- *Total access frequency/duration of visits longer than 2 hours*: the total times/duration of visits of the place longer than 2 hours. People usually stay at home, work place, shopping center for more than 2 hours.
- *Average access frequency/duration in weekdays*: the total times/duration of the place visited in weekdays divided by 5, i.e. average frequency/duration per day. Most of people need to work during the day time in weekdays.
- *Average access frequency/duration in weekend*: the total times/duration of the place visited in weekend divided by 2. Many people may have some entertainment activities during weekend, such as shopping, visiting friends, or doing sports.
- *Total access frequency/duration in the daytime*: the total times/duration of the place visited between 7:00am to 7:00pm. Daytime is the time for work in the weekdays.
- *Total access frequency/duration at the nighttime*: the total times/duration of the place visited between 7:00pm to 7:00am. Night time is another time for fun and relax. Note this is different from sleeping time that we will define next.
- *Total access frequency/duration in the sleeping time*: the total times/duration of the place visited between 12:00am to 6:00am. People usually sleep at home during this time period.

Besides the time/space-related factors described above, other behaviors of people at a particular place may also be helpful on the detection of place's semantic meaning. For example, a person who is at transportation place (such as railway station) may have a higher moving speed and a larger variation of accelerometer readings before or after a visit of that place. A person doing outdoor sports (like jogging or hiking) may have a considerable moving speed and obvious accelerometer readings during the visit. On the contrary, for some places such as user's home, work place and restaurant, we probably will not monitor obvious or large variations of accelerometer readings. Therefore, people's movement behavior is another important factor for semantic place prediction. The accelerometer tracing file (*accel.csv*) in MDC dataset provides an array of records containing the relative time and acceleration data[1]. With these information, it is easy to calculate the average and variation of acceleration readings of a particular user at a place. We define and use the following movement-related features from accelerometer samples:

- *Average acceleration during the visit*: the user's average accelerometer readings when he/she visits the place. If the user has considerable acceleration when he/she visits the place, he/she may be doing sport.
- *Variation of acceleration during the visit*: the variation of user's accelerometer readings when he/she visits the place. If the user has considerable variety of accelerations when he/she visits the place, he/she may be in a shopping center.
- *Variation of acceleration before the visit*: the variation of user's accelerometer readings before he/she visits the place. This feature could be used to distinct the locations related to transportation.
- *Variation of acceleration after the visit*: the variation of user's accelerometer readings after he/she visits the place. This feature could also be used to distinct the locations related to transportation.

Notice that if GPS data is available, average speed and maximum position variety could also be used.

Last, detecting user's social ties from their Bluetooth scan record (*bluetooth.csv*), call logs (*callog.csv*), and address book (*contacts.csv*) could also help with the detection of a friend's home and workplace. For example, friends may call each other often, share similar entries in their address book, or meet frequently (see each other in their Bluetooth scan). However, unfortunately in the MDC dateset, the anonymized data could not be used for such kind of social relationship detection due to the privacy protecting reason. This makes it hard to predicate a friend's places.

2.2 Prediction with Rule Based Method

Comparing with other place categories, home and work places are two of the most important places in people's life. These places normally have the largest *total access frequency* and are easier to detect. An obvious feature of home is: most people sleep at home. Solely using this single fact (based on the feature of *total access duration in the sleeping time*) we can detect people's home. The workplace detection is a little bit complicated. As we have already observed, most of people's work places should have high *average access frequency in weekdays* and high *total access frequency in daytime*. They should also have low *average access frequency in weekend* and low *total access frequency at nighttime*. Following these observations, we can make home and work place detection based on the following two simple rules:

[1]We would like to thank Dr. Sanjiv Nanda for pointing out an error of our understanding of the accelerometer readings in an earlier version of this paper.

- **Home**: comparing the *total access duration in the sleeping time* of each place ID for a user, we set the one with highest value as the user's home.
- **Work place**: among the places of each user with the top five *total access frequency*, find the places, whose *average access frequency in the weekdays* is larger than its *average access frequency in the weekend* and its *total access frequency in the daytime* is larger than its *total access frequency in the nighttime*. Among these selected places, we identify the place with the largest *total access frequency* as the user's work place.

We evaluate these simple prediction method for home and work places over the ground truth data (Set A). Table 1 shows the detailed classification accuracy of rule based method for home and work places. Our home and workplace rule based detection methods can achieve relatively high classification accuracy. Notice that not everyone has distinct home and work locations: some people work at home, some have no fixed work site, and others may not use their cell phones at home. Therefore, it is impossible to correctly identify home and work location for all users using these simple rules or even more complex techniques.

With the movement-related features, locations related to transportation and shopping center might be detected based on the following simple rules:

- **Place related to transportation**: if the place's *variation of acceleration during the visit* is larger than 500 and its *variation of acceleration before or after the visit* is larger than 100 and smaller than 1000, we set this place as the location related to transportation.
- **Shopping center**: if the place's *variation of acceleration during the visit* is larger than 1000, we set this place as the place for shopping center.

Notice that here we use a few hard thresholds which need to be carefully adjusted depending on the applied dataset. It is obvious that this part of prediction is not as accurate as the one for home and workplace prediction. However, it could be a possible complement to other prediction methods. For example, purely using the machine learning based methods (we will introduce next) may suffer from the lack of enough instances in the ground truth data, especially for certain types of places. Thus combining rule-based methods with machine learning methods may improve performance.

2.3 Prediction with Machine Learning Method

In rule based methods, we manually formulate rules based on regular patterns to distinguish some types of places, however not all categories of places have such obviously regular patterns. We further explore more intelligent ways to learn hidden patterns of all types of places. Naturally, machine learning methods become our choice. Machine leaning techniques have been widely used to discover behaviors and patterns based on large-scale empirical data. Machine learning algorithms can take advantages of examples (training data) to capture characteristics of interest of the unknown underlying probability distribution. They could automatically learn to recognize complex patterns and make intelligent decisions based on data. For this project, we directly use WEKA [6], a comprehensive tool for machine learning and data mining, to explore the user's mobility pattern on different categories of places.

Table 1: Classification accuracy of our rule based method for home and work place prediction

Semantic Place	TP Rate	FP Rate
Home	0.762	0.081
None-Home	0.919	0.238
Workplace	0.765	0.105
None-Workplace	0.895	0.235

Table 2: Distribution of instances among categories

Place Label (Category)	1	2	3	4	5	6	7	8	9	10
Set A	84	46	102	23	9	25	14	11	17	5
Training set	64	35	70	15	6	20	10	8	12	4
Testing set	20	11	32	8	3	5	4	3	5	1

Basically, we define the semantic place predication as a classification problem. There are 10 classes representing different types of semantic places. We use the features introduced in Section 2.1 as the features for our classifiers. We use multiple well-known classifiers, including NaiveBayes, BayesNet, IBK, J48, AdaBoostM1. We first divide the ground truth data (Set A) into training data and testing data, and then train/evaluate the classifiers with these data. We randomly pick 3/4 of all instances in Set A as the training set and the remaining 1/4 instances as the testing set. Table 2 shows the distribution of these instances on different place categories.

We consider 10 individual classification tasks, each of which is for one place category and tests all five machine learning algorithms (classifiers) on them. Table 3 shows the truth positive rate and false positive rate of our methods for each place category and classifier respectively. Clearly, different machine learning algorithms have different predict accuracies especially for place labels with few instances.

Based on their classification accuracies, we either pick the best classifier or combine multiple classifiers to form different predication methods. To combine results from multiple classifiers, a simple majority vote could be used. Besides of building a classifier for each individual place category, we also build general classifiers to predict places among all 10 categories using different machine learning algorithms. We then combine these machine learning methods with rule based methods to generate integrated results. All these predication methods are trained and used in generating final submissions over the final testing MDC dataset (Set B).

3. TASK 2: PREDICTING NEXT PLACE

The second prediction task is to predict the next place given the current location of a user. We again leverage the power of machine learning approaches. Since user mobility is user dependent, we train a single classifier for each user to predict its next place. The place ID of the next destination is used as

Table 3: Classification accuracy (truth/false positive rate) of machine learning methods in Task 1

Place Label (Category)	Naive Bayes	Bayes Network	IBK	J48	AdaBoostM1
1	0.500 / 0.036	0.563 / 0.091	0.688 / 0.055	0.625 / 0.018	0.563 / 0.018
2	1.000 / 0.790	0.000 / 0.000	0.556 / 0.226	0.222 / 0.097	0.444 / 0.177
3	0.762 / 0.380	0.429 / 0.040	0.524 / 0.100	0.476 / 0.120	0.333 / 0.100
4	1.000 / 0.299	1.000 / 0.194	0.500 / 0.075	0.500 / 0.075	0.000 / 0.000
5	1.000 / 0.400	0.000 / 0.271	0.000 / 0.014	0.000 / 0.000	0.000 / 0.000
6	0.900 / 0.475	0.600 / 0.361	0.000 / 0.066	0.000 / 0.000	0.000 / 0.000
7	0.857 / 0.516	0.000 / 0.000	0.000 / 0.000	0.143 / 0.000	0.000 / 0.000
8	0.800 / 0.394	0.000 / 0.000	0.000 / 0.000	0.000 / 0.000	0.000 / 0.000
9	1.000 / 0.536	1.000 / 0.290	0.000 / 0.029	0.000 / 0.043	0.000 / 0.000
10	1.000 / 0.186	0.000 / 0.000	0.000 / 0.014	0.000 / 0.000	0.000 / 0.014

Table 4: Prediction accuracy (correct classification ratio) of machine learning method for 10 users in Task 2

User ID	25	43	49	59	81	99	107	146	162	177
Prediction accuracy (IBK)	0.417	0.66	0.574	0.469	0.723	0.454	0.377	0.489	0.652	0.583

the target or class variable for prediction. The key question is how to choose features for such classification task.

Since the next place heavily depends on the current location of the user, it is obviously that both spacial and temporal information of current context could be helpful for the prediction task. We extract both types of features from the sequence of visited place (*visit_sequence_20min.csv*) to train and test our machine learning based prediction methods.

- *Place ID of the current context*: the place that this user was leaving. It is clear that the place sequence should be continues over space.
- *Duration of the current context*: how long the user stayed at this place may also affect the movement to the next place.
- *Day of the week of the end time of current context*: which day in a week of current time. We believe people may have different mobility patterns in weekdays and weekend, or even different for each day within a week.
- *Hour of the day of the end time of current context*: which hour in a day of current time. People's access regulation is also related to the time of the day.

For each user, we train separate models using these features over the training set (Set A) built from their own *visit_sequence_20min.csv*. Since the movement between places should be continuous, we only consider those instances whose *trusted_transition* is 1. We test our methods over the ground truth data *toy_nextplace_segment.csv* for 10 different users. The prediction accuracies with underlying classifier of IBK are listed in Table 4. Clearly, the prediction of next place is a challenging task. For the final submissions to MDC contest, we apply these trained models on the testing set (relative to the time interval for each test data point and features are extracted from Set C and *nextplace_segment.csv*). We again try different classifiers and combine their results in our final submissions.

4. CONCLUSIONS

In this paper, we introduced our proposed methods to predict the semantic meaning of the "important places" and the users' next destination in Nokia MDC dataset. We have explored different temporal and spacial features of the visited place logs and accelerometer samples. Two types of prediction methods are proposed: rule based and machine learning based. Preliminary results over released MDC dataset show that the proposed methods can achieve reasonable prediction accuracy if the number of instances in the training data is sufficient.

Besides the methods reported here, we tested other approaches and techniques, such as feature normalization and feature selection, however, the improvement is not significant. We also tried to further train two sub-types of work places (one with longer duration and the other with shorter duration), but the overall accuracy does not change. For Task 2, we would also consider the average speed or relevant distance among places to predict the next place. However, without accurate time and distance, such approach does not work. We leave further study of other types of methods or features for both tasks to improve the predication accuracy as our future works. Additionally, we plan to apply the discovery from this study to help with the design of new network protocols for pocket switched networks [7, 15] or delay tolerant networks [18].

Finally, we would like to thank the Nokia MDC organizers to provide such a great opportunity for us to participate in this challenge. We hope that Nokia research and other cellular companies can further release more high quality dataset to research community.

5. REFERENCES

[1] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, pages 275–286, 2003.

[2] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of

one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10:18–26, 2011.

[3] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 2012.

[4] N. Eagle, A. Pentland, and D. Lazer. Inferring Friendship Network Structure by Using Mobile Phone Data. *PNAS*, 106(36), 2009.

[5] S. Eubank, H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling disease outbreaks in realistic urban social networks. *Nature*, 2004.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[7] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *WDTN '05: Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251, New York, NY, USA, 2005. ACM.

[8] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th international conference on Pervasive computing*, Pervasive'11, pages 133–151, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, WMASH '04, pages 110–118, New York, NY, USA, 2004. ACM.

[10] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge 2012 (by Nokia) Workshop, in conjunction with Int. Conf.. on Pervasive Computing*, Newcastle, June 2012.

[11] N. Marmasse and C. Schmandt. A user-centered location model. *Personal Ubiquitous Comput.*, 6:318–321, December 2002.

[12] D. J. Patterson, L. Liao, D. Fox, and H. Kautz. Inferring high-level behavior from low-level sensors. In *Ubicomp*, volume 2864, pages 73–89, 2003.

[13] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, July 2007.

[14] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie. Mining individual life pattern based on location history. In *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, MDM '09, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society.

[15] L. Zhao, F. Li, C. Zhang, and Y. Wang. Routing with multi-level social groups in mobile opportunistic networks. In *Proc. of IEEE Globecom*, 2012.

[16] V. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proc. of WWW' 10*, pages 1029–1038, 2010.

[17] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS tragectories. In *Proc. of WWW' 09*, pages 791–800, 2009.

[18] Y. Zhu, B. Xu, X. Shi, and Y. Wang. A survey of social-based routing in delay tolerant networks: Positive and negative social effects. to appear in IEEE Communication Survey and Tutorials, 2012.