

Mobile Users Location Prediction with Complex Behavior Understanding

Mostafa Karimzadeh, Zhongliang Zhao, Florian Gerber, Torsten Braun

Institute of Computer Science, University of Bern, Switzerland

Email : {karimzadeh, zhao, braun}@inf.unibe.ch, florian.gerber@students.unibe.ch

Abstract—The growing ubiquity of smart-phones equipped with built-in sensors and global positioning system (GPS) has resulted in the collection of large volumes of mobility data without the need of any additional devices. The large size of heterogeneous mobility data gives rise to rapid development of location-based services (LBSs). The predictability of mobile users' behavior is essential to enhance LBSs. To predict human mobility, many techniques have been proposed. However, existing techniques require good data quality to guarantee optimal performance. In this paper, we proposed a hybrid Markov chain to predict mobile users' future locations. Our model constantly adapts to available user trace quality to select either the first order or the second order Markov chain. Compared to existing solutions, our model is adaptive to discrete gaps in data trace. To help us understanding complex user behaviors, we have also proposed a technique benefiting both temporal and spatial parameters to extract Zone of Interests (ZOIs). To evaluate the algorithm's performance, we use a real-life dataset from the Nokia Mobile Data Challenge (MDC) collected around Lake Geneva region from 180 users. We found a satisfactory future user location prediction accuracy of 70 – 84%.

Index Terms—Mobile analysis, Mobility Prediction, Mobility Behavior, Location based Services.

I. INTRODUCTION

Extracting meaningful information from collected trace data of users to determine their movement pattern is an important part of location-based services (LBSs). For example, to predict future behavior of a mobile user, mobility predictors rely on clustering techniques to capture a user's Individual Zone of Interests (I-ZOIs) from collected trace data. Intuitively, an I-ZOI is a city area that is frequently visited by a user and the user spends considerable time in this region. Typically, LBSs are using mobility prediction as a means to improve quality of service by providing context-aware information to users beforehand.

In recent years, we have seen a rapid proliferation in the number of services such as *Google Now*, which proactively collects rich contextual data, such as Bluetooth/WiFi connectivity, call/SMS log, information about running applications, to deliver information to users that they may need in daily life activities. Similarly, *Google Maps* is a web mapping service to predict future locations of users based on the movement history. It is apparent from the above examples that location based services are thriving, providing a remarkable opportunity to collect fine grained data about visited places of users. This source of user mobility offers new possibilities to tackle established research problems on human mobility.

In this work, the fundamental goal is to formulate the location prediction problem using a Mobility Markov Chain model (MMC). We propose a dynamic Markov chain-based model, which adaptively selects the first-order or the second-order Markov chain model based on the available trace quality. We propose a clustering algorithm to discover frequently visited places by the user and integrate it with a mobility predictor to predict a user's future behavior.

The remainder of the paper is organized as follows. Section II discusses research efforts. Section III presents our system model and introduces some formal definitions and notations used in the paper. Sections IV and V discuss the methodology to evaluate the mobility model and demonstrate obtained results respectively. Finally, Section VI concludes the paper by sketching future research directions.

II. RELATED WORK

Understanding human mobility by mining raw GPS logs has been a long-standing subject in academic research [1], [2]. These approaches rely on clustering user visits to extract hot spots. The premier research in adapting the data clustering algorithms for modeling mobility behavior [3] proposes to iteratively extract hot spots of users. Montoliu et al. proposed a clustering algorithm [4], where GPS coordinates (*latitude and longitude*) are clustered in the temporal domain to detect the stay points that are used to derive frequently visited regions using a grid-based clustering approach.

The above techniques use several temporal bounds to classify a particular region as a cluster. Some parameters include maximum distance between the collected locations, maximum and minimum time bound, cluster shapes. However, if we only take into account the temporal bounds, this leads to some inaccuracies in estimating the total number of clusters belonging to a user. As opposed to using only temporal metrics to cluster the individual regions, we form the clustering algorithm by benefiting both temporal and spatial metrics (*instantaneous velocity, average velocity*) to quantify the correlation between visited regions.

Regarding the prediction techniques, a majority of existing works first formulate a mobility model and consequently use it to make prediction [5],[6]. Numerous map-matching algorithms—that matches a sequence of location samples to road network—have been proposed to predict user mobility. However, most existing map-matching techniques are not always practicable and need additional services such as network

connectivity [7]. Such schemes completely ignore the trace data with poor quality and just consider users with good quality of trace data. Our work consists of estimating the frequently visited locations, where a user spends considerable amount of time, and then attaining the hybrid Markov chain model, which adaptively selects from the first-order or the second order Markov chain, depending on the quality of user traces to predict future behavior of the users.

III. SYSTEM MODEL

The system overview is depicted in Figure 1. The proposed system model involves two main layers: Individual Zone of Interest (*I-ZOI*) Discovery and Individual Zone of Interest (*I-ZOI*) Prediction. The relevant notations and system component definitions are explained in the following subsections.

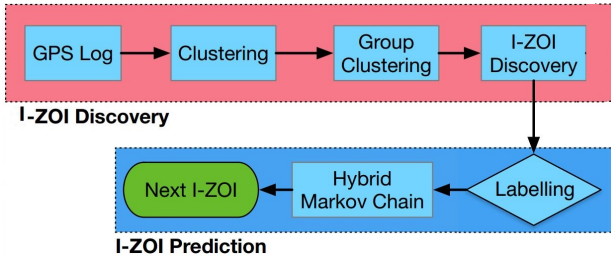


Fig. 1: Overview of the System Model.

A. Locations and Places

Mobile devices have the potential to track users' visited locations using the Global Positioning System (GPS). The device regularly collects a user's raw location logs as a list $L = [loc_1, loc_2, loc_3, \dots, loc_n]$, where $loc_i = (\alpha, \beta, t, v)$ is a tuple representing a location point in the format (latitude, longitude, timestamp, velocity). The rest of the paper uses the notations $loc.\alpha$, $loc.\beta$, $loc.t$ and $loc.v$ for the location point elements. In addition to GPS coordinates, users' daily activities consist of some places that they find useful or where they spend a considerable amount of time. This work focuses on places, which we refer to Zone of Interest (ZOI) hereafter.

B. Individual Zone of Interest (C-ZOI) Discovery

Intuitively, an Individual Zone of Interest (*I-ZOI*) is a cluster group that is frequently visited by a user. In order to extract the *I-ZOIs* of users based on the location history L , we must first introduce the notations of cluster and cluster group. A cluster represents a hot spot in an encapsulated area. It includes a subset of locations with similar temporal (e.g., visiting time, visitation repeatability rate) and spatial (e.g., speed, acceleration) features. A cluster group is an aggregation of intersected clusters.

1) *Cluster Discovery*: Location points with common temporal and spatial characteristics are considered as a cluster. Δd_{max} , e_{max} , e_{min} and $v \in \mathbb{R}$ represent the distance,

maximum velocity, minimum velocity and instantaneous velocity, respectively. In addition, $\Delta t_{min} \in \mathbb{N}$ is a time duration expressed in minutes. We introduce two functions: $ClusterCentroid([loc_1, loc_2, loc_3, \dots, loc_n])$, to compute the centroid of visited location points, and $Distance([loc_i, loc_j])$, which measures the Euclidean distance between the two locations loc_i and loc_j . A subset $l \subseteq L$ becomes a cluster if the following conditions in Equation 1, 2, 3 and 4 are met:

$\forall loc_i, loc_{i+1} \in l :$

$$Distance(centroid(loc_1, \dots, loc_i), loc_{i+1}) \leq \Delta d_{max} \quad (1)$$

$$loc_n.t - loc_1.t \geq \Delta t_{min} \quad (2)$$

$$e_{min} \leq \sum_{i=1}^n \frac{v_i}{n} \leq e_{max} \quad (3)$$

$$\nexists l' \neq : l \subset l' \quad (4)$$

A cluster is a 4-item tuple $c = (\alpha_c, \beta_c, \Delta r, l)$, where α_c and $\beta_c \in \mathbb{R}$ are the latitude and longitude coordinates of the centroid, $\Delta r \in \mathbb{R}$ is its radius, and $l \in L$ is the subset of locations belonging to c . The average of all $loc.\alpha$ and $loc.\beta$ of the locations contained in the subset l is the centroid (α_c, β_c) of the cluster, which is designated as $c.centroid$. We introduce C , the set of clusters extracted from the location log of a user as $C = \{c_1, c_2, c_3, \dots, c_n\}$. Disjointness of discovered clusters can not be guaranteed by equations 1, 2, 3 and 4 that is in turn to explain cluster group construction in next step.

2) *Cluster Group Discovery*: A cluster group includes a set of overlapped clusters. Thus, we define equation 5 to check whether two clusters $c_i, c_j \in C$ are intersected or not.

$$Distance(c_i.centroid, c_j.centroid) - (c_i.\Delta r + c_j.\Delta r) < 0 \quad (5)$$

A cluster group is a 4-item tuple $cg = (\alpha_{cg}, \beta_{cg}, \Delta r, \{c_1, c_2, c_3, \dots\})$, where α_{cg} , β_{cg} and $\Delta r \in \mathbb{R}$, $\{c_1, c_2, c_3, \dots\} \in C$ are latitude, longitude, radius and array of clusters constituting g respectively. $(\alpha_{cg}, \beta_{cg})$ represents the centroid of the cluster group, which is the mean of all the clusters formed g , and Δr must be compared to enclose all the individual clusters present in g . G contains the n cluster groups belonging to a user as $G = \{cg_1, cg_2, cg_3, \dots, cg_n\}$.

3) *Individual Zone of Interest (I-ZOI)*: An *I-ZOI* refers to a visited region by a user frequently and during daily activities. We define two constants $minCountThreshold$ and $maxTimeDifference \in \mathbb{N}$ representing the minimum threshold of visits and the maximum time difference threshold between two consecutive visits, respectively. Then, $CountVisits(cg)$ is a function that counts the number of clusters included in cluster group cg , and $timeDifference(G)$ is a function that returns the duration between two consecutive visited dates of cluster group cg in G . A cluster group $cg \in G$ is transformed into an *I-ZOI* z if the conditions of Equation 6 and 7 are met:

$$CountVisits(cg) \geq minCountThreshold \quad (6)$$

$$\text{timeDifference}(G) \leq \text{maxTimeDifference} \quad (7)$$

An I-ZOI z is a 6-item tuple $z = (\alpha_z, \beta_z, \Delta r, ID_{zone}, \{g_1, g_2, g_3, \dots, g_n\}, T_{ID})$, where α_z, β_z and $\Delta r \in \mathbb{R}$, $ID_{zone} \in \mathbb{N}$ and $\{g_1, g_2, g_3, \dots, g_n\} \in G$ are the latitude, longitude, radius, Zone-ID and group clusters becoming an I-ZOI. T_{ID} represents visiting dates of the I-ZOI by each user. The set Z is finally the set of I-ZOIs of the user such that $Z = \{z_1, z_2, z_3, \dots, z_n\}$. As shown in Figure 2, the set of discovered cluster groups are depicted by intersected yellow circles. Finally, the cluster groups that could satisfy above conditions are considered as I-ZOIs.

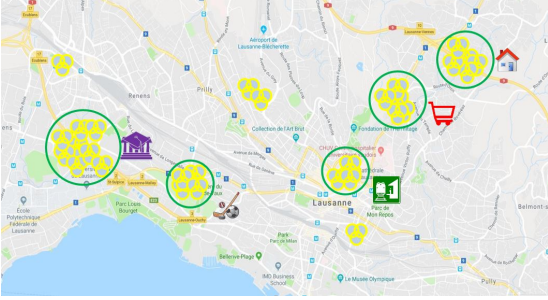


Fig. 2: I-ZOI construction from cluster groups of location points.

$$\begin{aligned} \text{Distance}(I - ZOI_i.\text{centroid}, I - ZOI_j.\text{centroid}) \\ \leq \text{distanceThresold} \end{aligned} \quad (8)$$

C. Individual Zone of Interest (I-ZOI) Prediction

The proposed hybrid Markov model benefits from both the first-order Markov chain [8] and the second-order Markov chain. The rationale behind using a hybrid predictor is that the standard first-order Markov chain algorithms are memoryless models, which means that the input for the next-place prediction task includes current visited location, current time and the day of week that the user is in the movement. The second-order Markov chain model is slightly different. In addition to current state, it benefits from previous state to predict future location.

Equation 9 defines proposed hybrid model, in which Z_i represents an I-ZOI with ID i , D indicates the day of the week (e.g., Saturday), T_i defines the time of the day D (e.g., 13:22:43 h), and λ_p determines the future time interval to predict next movement of the user.

$$Pr(Z_{i+1}(t + \lambda_p)) = \begin{cases} Pr(Z_{i+1}|Z(t + \lambda_p) = Z_i, \\ D, T(t + \lambda_p) = T_i) & Z_{i-1} = 0 \\ Pr(Z_{i+1}|Z(t) = Z_{i-1}, Z(t + \lambda_p) = Z_i, \\ T(t) = T_{i-1}, T(t + \lambda_p) = T_i, D) & Z_{i-1} \neq 0 \end{cases} \quad (9)$$

IV. EVALUATION

In this section, we present an evaluation methodology to validate the proposed user mobility prediction model.

1) *Dataset*: We used a mobility data trace collected during the Nokia Mobile Data Challenge (MDC) [9]. This dataset includes rich context information from the mobile phones for around 180 users around the lake Geneva region in Switzerland from October 2009 to March 2011. It includes Global Positioning System (GPS) information, running applications, chat records, calling records, etc. However, for mobility prediction, we are only interested in GPS location information, which are more than 10 million location points. For each user, we separated available data into two parts: the first part is the learning data set (L) and the rest is the testing data set (T). The learning data set includes the first 70% of user data. It is used to drive the states for both algorithms and to determine their transition probability matrix. The testing data set T contains the last 30% of the traces, which is used to test and evaluate the accuracy of the proposed prediction algorithm.

2) *User Trace Quality*: Different behaviors of users lead to different trace qualities. Some users carry the smart-phone all the time. However, some others forgot to carry the devices or had to charge them, such that data recordings are non-continuous. As learned from our previous experiences [10] [11], the number of valid states (with a time stamp and I-ZOI ID) in the driven hybrid Markov chain for each user depend on the quality of data trace in each day. Therefore, we first classify the dataset into two groups (good or poor quality) based on the number of recorded instances during the whole data collection period. We choose five users with good quality of trace data (e.g., 500000-400000 records) and five users with poor quality of trace data (e.g., 250000-350000 records).

3) *Evaluation Metrics*: Prediction accuracy measures the accuracy of the location prediction algorithm. We randomly select 10% of the states out of all the Markov chain states (e.g., states from 9 AM to 11 AM) derived for each particular weekday from the training data set L for each user. Afterwards, the prediction algorithm is performed for each of the selected states to estimate the possible future visited I-ZOI(s) for mobility prediction in the next λ_p minutes. These states have been chosen as random testing points. We check the transition probability for states during the same period of time in the testing data set T as well. Afterwards, the Mean Absolute Error (MAE) of the possible transitions of the corresponding testing points is calculated according to Equation 10.

$$MAE = \frac{1}{N} \sum_{i=1}^N |Pr_i L - Pr_i T|, \text{Accuracy} = (1 - MAE) \times 100 \quad (10)$$

4) *Experimental Settings*: We describe the experimentation parameters of the discussed clustering and mobility prediction algorithms. In order to determine the parameters we analyze traced data for users with at least 10 months duration of collected data. Then, we read the data-points sequentially according to the recorded time stamps. Table I shows the experiment parameters and the associated values in our assessment.

V. EVALUATION RESULTS

1) *Mobility Prediction Accuracy Results*: This subsection details the prediction accuracy results of the proposed hybrid

TABLE I: Experiments parameters.

Parameter	Definition	Value
Δd_{max}	Maximum Euclidean distance between centroid of location points and next location	60 m
Δt_{min}	Minimum time threshold of staying in each location point	15 min
e_{max}	Maximum instantaneous velocity threshold	50 m/s
e_{min}	Minimum instantaneous velocity threshold	0
minCountTh	Minimum number of visits of each cluster group	60
maxTimeDiff	Maximum time difference between two consecutive visits of a cluster group by a user	24 h
DistanceTh	Maximum Euclidean distance between two I-ZOIs	500 m
λ_p	Time threshold for hybrid-MC algorithm	1 min

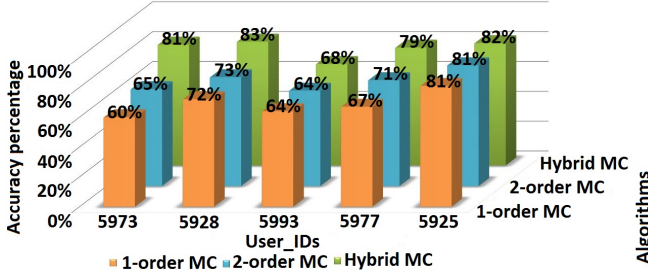


Fig. 3: Prediction accuracy for users with good quality.

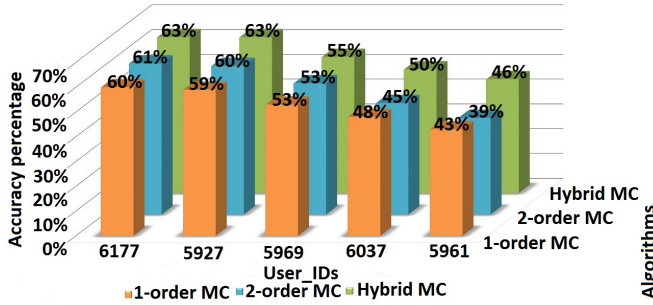


Fig. 4: Prediction accuracy for users with poor quality.

predictor and the first order and the second order Markov chain.

Figures 3 and 4 show the prediction accuracy of different MMC predictors for users with good and poor quality of mobility traces. We define two categories of quality depending on the number of instances recorded in a user's movement traces. We randomly choose 5 User IDs (5973, 5928, 5993, 5977, 5925) from the group of good quality trace data, and 5 User IDs (6177, 5927, 5969, 6037, 5961) from the group of poor quality trace data. As we can see from Figure 3, the hybrid predictor can deliver an average prediction accuracy over all weekdays and weekends of nearly 83% for User-ID 5928. Moreover, it can be observed from Figure 4 that the estimated accuracy is improved significantly when the hybrid predictor used for users with poor quality trace data. For instance, it delivers an average accuracy of 63% for User-IDs 6177 and 5927. The results clearly demonstrate that the hybrid predictor outperform others, while using the traced data with either poor or good quality.

VI. CONCLUSIONS

With the explosive growth of location-based services on mobile devices, predicting users' future locations is of increasing importance to support proactive information services. In this paper, we introduce a hybrid predictor to estimate future locations of a user. Further, we propose a technique to discover hot spot regions for users by relying on spatial and temporal constraints. The achieved results over real world mobility traces validates our proposed algorithms, which achieve more than 81% correct predictions for users. For future improvements we will focus on predictability of users congestion in city areas.

ACKNOWLEDGMENT

This work has been supported by the Swiss National Science Foundation with project number 154458.

REFERENCES

- [1] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles," *ACM Trans. Web*, vol. 8, no. 4, pp. 21:1–21:25, Nov. 2014. [Online]. Available: <http://doi.acm.org/10.1145/2637483>
- [2] V. Kulkarni, A. Moro, and B. Garbinato, "Mobidict: A mobility prediction system leveraging realtime location data streams," in *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming*, ser. IWGS '16. New York, NY, USA: ACM, 2016, pp. 8:1–8:10. [Online]. Available: <http://doi.acm.org/10.1145/3003421.3003424>
- [3] T. N. Maeda, K. Tsubouchi, and F. Toriumi, "Next place prediction in unfamiliar places considering contextual factors," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL'17. ACM, 2017, pp. 76:1–76:4.
- [4] 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*, ser. MUM '10. ACM, 2010, pp. 12:1–12:10.
- [5] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '11. New York, NY, USA: ACM, 2011, pp. 34–43. [Online]. Available: <http://doi.acm.org/10.1145/2093973.2093980>
- [6] S. Gams, M.-O. Killijian, and M. N. n. del Prado Cortez, "Next place prediction using mobility markov chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, ser. MPM '12. New York, NY, USA: ACM, 2012, pp. 3:1–3:6. [Online]. Available: <http://doi.acm.org/10.1145/2181196.2181199>
- [7] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, p. 2923, Oct. 2013.
- [8] M. Karimzadeh Motallebi Azar, Z. Zhao, L. Hendriks, R. de Oliveira Schmidt, S. la Fleur, H. van den Berg, A. Pras, T. Braun, and M. Julian Corici, *Mobility and Bandwidth Prediction as a Service in Virtualized LTE Systems*. IEEE, 10 2015, pp. 132–138, 10.1109/Cloud-Net.2015.7335295.
- [9] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data," *Int. J. Inf. Manag.*, vol. 36, no. 6, pp. 1231–1247, Dec. 2016. [Online]. Available: <https://doi.org/10.1016/j.ijinfomgt.2016.07.009>
- [10] Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, and S. Sargento, "Mobility prediction-assisted over-the-top edge prefetching for hierarchical vanets," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2018.
- [11] Z. Zhao, M. Karimzadeh, F. Gerber, and T. Braun, "Mobile crowd location prediction with hybrid features using ensemble learning," *Future Generation Computer Systems*, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167739X17318058>