2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing

Predicting Mobile Phone User Locations by Exploiting Collective Behavioral Patterns

Haoyi Xiong, Daqing Zhang, Daqiang Zhang and Vincent Gauthier

Institut Mines-Télécom, Telecom SudParis, CNRS SAMOVAR, Evry, 91000, France
{*haoyi.xiong, daqing.zhang, daqiang.zhang, vincent.gauthier*}*@telecom-sudparis.eu*

***Abstract*—Location prediction based on cellular network traces has recently spurred lots of interest. However, predicting one's location remains a very challenging task due to the randomness of the human mobility patterns. Our preliminary study included in this paper shows that there is a strong correlation and association among the certain group of users' locations. Through association pattern mining on Reality Mining dataset which involves 32,579 cell tower locations and 350,000 hours of continuous activity information, we observe the highly confident association rules exist among the locations of users, and then we further verify that the associations are indeed caused by the *collective behaviors* of the mobile phone users.**

Based on this finding we introduce the *collective behavioral patterns* (CBP), and then propose CBP-based predictor— a novel prediction schema that aims to forecasting one's locations in next 6 hours based on the locations of other users. Furthermore, we integrate the state-of-the-art i.e., Markov-based predictor with our CBP-based schema to build a hybrid predictor. We evaluate the CBP-based schema and compare the hybrid predictor with the Markov-based predictor through intensive experiments. Experimental results show that CBP-based predictor achieves good precision and the hybrid predictor produces higher prediction accuracy than the state-of-the-art scheme at cell tower level in the forthcoming one to six hours. Finally it is verified that *collective behavioral patterns* can be used to predict user locations as well as to improve the performance of existing predictors.

I. Introduction

Predicting mobile phone users locations in next few hours is essential for a wide range of ubiquitous computing applications including location-based services [1], and resource management of telecommunication network [2]. However, it remains challenging to foretell one's locations owing to the high degree of freedom and individualism of user mobility patterns [3].

The goal of this research is to predict the locations of mobile phone users in next few hours. Particularly, we are interested in improving the precision of existing schemas like Markov-based predictor leveraging the *collective behaviors*. The *collective behaviors* hereby refer to the behaviors of crowds emerging spontaneously. While we agree that the main driver of individual's movement is the regularity of her own mobility, however we also find that the *collective behaviors* affect individual's mobility. Therefore it's logical to recognize and use *collective behaviors* for location prediction. Existing approaches for human location prediction are many. They mainly focus on the temporal-spatial regularity of individual's mobility, for instance, the Markov-based predictor [4]. However the predictability of individual's mobility pattern seems theoretically limited [3]. Some pioneering work [5], [6] has observed or even has measured some social factors affecting individual's mobility. Thus far, the social factors concerned for the most are the social contacts and interactions [5]. However as a kind of social factors, *collective behaviors* don't rely on the implicit social interactions but much more emphasizes the crowds' behaviors that are occurred "spontaneously". Unfortunately, to the best of our knowledge, few work [6] has been done to bring the observation and measurement of *collective behaviors* into designing of location predictor.

In this work we uncover the *collective behavioral patterns* (**CBP**) – i.e., the association patterns [7] among the locations of mobile phone users. The *collective behavioral patterns* are named after our observation in the realistic dataset that the user locations of the same type e.g., residences or offices are frequently associated. For example when 90% users are staying at their own residences, then the rest 10% are probably located in their residences as well, although these residences are usually in the different locations. It is due to the collective behaviors, which means the crowds may stay at the same type of location spontaneously. We observe thousands of cases of the *collective behaviors* through mining the association patterns from user mobility data of Reality mining dataset [8]. The observed association patterns are exactly the *collective behavioral patterns* we mentioned in this paper. For the detailed observations, we will address them in section IV. Thus our work tries to adopt the *collective behavioral patterns* to predict the user's location in next hours from the current locations of other users. Through our predictor design and evaluation, our **CBP**-based schema achieves $40\% - 50\%$ precision alone.

Please note that the location prediction in our research is in the cell tower level; and each cell tower ID identifies a specific location. Foremost, being able to forecast user location in the cell tower level would benefit resource allocation, service handover and delay-tolerant routing to

978-0-7695-4843-2/12 $26.00 © 2012 IEEE
DOI 10.1109/UIC-ATC.2012.28

pervasive urban planning and intelligent traffic engineering. We conduct our research on Reality dataset [8]. It records the 1,000,000 GSM traces, 112,508 cellular calls and 350,000 hours of continuous human activity traces of 106 users, including staffs, professors and students, in MIT Campus–an urban area. Since the size of a cell varies from tens of meters in urban environment to kilometers in rural regions, the location positioning and predicting in our research is fine-grained.

Here goes the introduction to three main parts of our work.

- Firstly we observe the existence of *collective behavioral patterns* through association pattern mining, and uncover the association rules of **CBP**. Our empirical study shows it is confident to infer individual's locations from the locations of crowds.
- The association rules identify the ***correlation of user's locations at the same moment***; however, actually, the location prediction acquires ***correlations from the current locations of crowds to the locations of target user in next few hours***. We need to extend the *collective behavioral patterns* to associate users' locations from crowds to individuals with time-shifting. Hereby, we design a **CBP**-based Bayesian model to learn the correlations with time-shifting from the mobility data of crowds, and finally we enable this model into a location predictor.
- Since the design of our **CBP**-based predictor doesn't take the individual's mobility pattern into consideration, we propose a fusion mechanism to integrate the **CBP**-based predictor with the state-of-the-art of the individual mobility prediction schema i.e., Markov-based predictor [4]. The hybrid predictor achieves a prediction accuracy of 5%-10% higher than the Markov-based predictor.

To summarize, the main contributions of this paper are three folds.

1) To the best of our knowledge inspired by association rule mining result we are the first to model the *collective behaviors* in user mobility.
2) We try to adopt *collective behavioral patterns* in mobility predictor design. It shows that our **CBP**-based predictor is quite accurate.
3) The experimental result shows the performance of existing approaches based on individual mobility patterns is improved by integrating with our **CBP**-based predictor.

The rest of this paper is organized as follows. Section II briefly overviews the related work. Section III presents the problem statements and preliminary data pre-processing for this work. Section IV conducts empirical study on *collective behavioral patterns* measurement. Section V discusses our approach to model the *collective behavioral patterns* into location prediction as well as the design of **CBP**-based prediction schema. A hybrid predictor based on both **CBP**-based and Markov-based schemas is addressed in section V. Section VI reports the experimental results. Section VII concludes the work.

II. Related Work

A variety of schemes that address the problem of prediction of user location have been studied. In general, they fall into the schemas based on individual's mobility patterns, the schemas based on social-ties, and hybrid schemes integrating above two.

A. The schemes based on the individual's mobility patterns or based on social ties

These schemes take advantage of the temporal and spatial regularities that are exhibited in the individual's mobility patterns. The prediction schemas based on *markov models*, especially those based on the *higher-order markovian model* [4] are considered as the state-of-the-art in the practical predictor design [9], since it takes the probable locations for next movement and the temporal order of movements into account. Besides, some of other schemas foresee user location by detecting periodic patterns in user traces. The predictability of prediction schemes based on individual's mobility patterns is limited, around 90% in the theoretical upper bound [3].

They postulate that user movement is driven by social-tie, involving the social community identification, and the prediction based on the community attraction to users. However, social-tie is an elementary building block for user mobility, but not the only driver [10]. As a typical example, CMM [11] leveraged user friendship to cluster users as communities, and then decided user next location by community attraction.

B. The schemes integrating individual's mobility pattern with social-ties

In recent years, many hybrid schemes on predicting user location have been studied. Calabrese *et al.* [6] introduced the first realistic predictor fusing the *collective behaviors* and individual mobility patterns for mobile phone users. It employs a prediction schema based on the periodicity of individual's mobility pattern, and then uses the collective geographical preferences to refine the prediction result. Comparing with our research, instead of the collective preference, we choose another pointcut i.e., the association properties to model the *collective behaviors.* HCMM [12] was a mixture mobility model that closely resembled CMM. It fused location preference and social attraction together into prediction. Many other pionnering schemas like [13] will not be introduced here.

III. Data Pre-processing and Problem Statement

In this section we present the ping-pong effects in realistic mobile phone cell-ID logs. Then to simplify our research,

we pre-process our mobility data so as to reduce the noise caused by ping-pong effects. Finally we formulate the problem with respects to the characters of cellular-based mobility data.

A. Noise and Data Pre-processing

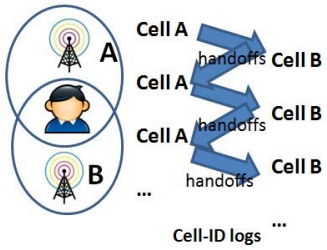


Figure 1: Noise of Ping-pong Effects

This section presents the typical noise in cellular-based location data and tries to uncover the reasons such noise caused by. Finally, we briefly introduce our data pre-processing stage to reduce the ping-pong noise.

Cell-ID logs are a sequence of user locations with time stamped and represent the trajectories of users' mobility. Mobile phone would make a log once user handover to another cell tower which may have better quality of service (e.g., strength of signal). However, when user stays in the overlapping coverage of multiple cellular towers without any movement, the mobile phone may frequently handoff among nearby cell towers even "when radio link still acceptable" [14]. Figure 1 illustrates a basic scenario that one mobile phone user stays under the coverage of two cell tower. The frequent handoffs between cell A and B, even the user has no movement at all. These handoffs are considered unnecessary handoffs, and the frequent switch with cells of surroundings caused by unnecessary handoffs is named as ping pong effects [14].

To filter out the noise data, our work maps all cells into non-overlapping regions, and identifies each region with the full set of cell towers which cover the region. For example, in Figure 2, region $\{A, B\}$ is assigned to the location covered by cell tower A and B but is out the coverage of cell tower C; the region $\{A, B, C\}$ identifies the location covered by cell tower A, B and C; furthermore the region $\{A, B\}$ and $\{A, B, C\}$ has no overlap. Finally we adopt single-reference area measurement [15] to localize users' region from their cell logs. Totally, we generate 34546 regions from 32579 cell towers in Reality Mining dataset. In our research, the locations for inputs and output of our predictors are regions.

B. Problem Statement

Through our data pre-processing, a user's cell log is converted into a list of regions with time stamped. We pick up the longest stay region for each hour i.e., the region where user spent the longest time for every hour slot, and then thread these regions into a sequence. Such sequence represents the trajectory of user's mobility.

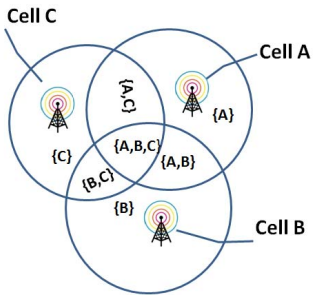


Figure 2: Region Extractions

Table I: Statistics of MIT Reality Mining Dataset

Item	Description	Item	Description
Starting time	Jan/2004	Ending time	Jul/2005
# of users	106	# of faculties	11
# of cell towers	32,579	# of areas	1,027
# of GSM trace	2,667,895	Avg. # of trace	46.7
# of mobile calls	112,508	Avg. # of call	≈ 4
Logical location	lac.cell	Physical location	no

Definition 1. A user's trajectory is a sequence of regions that user spend the longest time for each time slot, i.e.,

$$t : r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_n$$

where r_i identifies the longest stay region in i_{th} time slot.

Since one of our research goals is to exploit associations or exactly, the *collective behavioral patterns* to predict user locations, our problem formulation naturally relies on the locations and trajectories for a set of users rather than individual's location.

Problem. Given a set of users $U = \{u_1, u_2, \ldots, u_n\}$ and the set of trajectories for these users $T = \{t^1, t^2 \ldots t^n\}$ where $t^i = r_1^i \rightarrow r_2^i \cdots \rightarrow r_m^i$ (r_m^i denotes u_i's location in the m_{th} time slot); predict $r_{m+\tau}^k$, i.e., the region of user u_k being about to stay at the future time τ where $1 \leq k \leq n$ and $\tau \in \mathbf{N+}$.

In latter part of this paper, by exploiting the association properties of mobile phone user's locations, we propose a predictor to forecast one user's locations from other users' locations for next 6 hours (i.e., $\tau \in (0, 6]$).

IV. Measuring Collective Behavioral Patterns in Human Mobility

In this section, we aim at validating the existence of *collective behavioral patterns* (**CBP**) and measuring the effects of **CBP** to human mobility. The final goal of this empirical study is to investigate ***if it is possible to infer one user's location from the locations of other users***. To achieve the goal, we try to discover the *association patterns and rules* in user's locations of mobility.

A. Empirical Study Setup

This empirical study was made on MIT Reality Mining dataset [8]. Table I addresses a brief overview to MIT Reality Mining dataset. Due to the limited of dataset, for example many users are inactive sometimes, we simply pick up 15 active users [1] from all 106 users as well as their cell logs dating from 01-Nov-2004 to 29-Dec-2004. All cell logs are converted into the sequences of regions.

The association patterns that we attempt discover are based on the locations of users. Therefore each of the items is a user specified region, i.e., the tuple $\langle r_a, u_i \rangle$ which represents the user u_i staying in region r_a. We have found 1868 tuples for these 15 users. Therefore the transaction [7], i.e., locations of users in the same time slot, is specified as a set of tuples. More specifically, in this empirical study, each transaction is a set of 15 tuples.

Following above specifications, an association rule becomes the inference from left-hand set (**LHS**) of the tuples to right-hand set (**RHS**) of the tuples, for example:

$$\{\langle u_i,\, r_a\rangle\, ,\langle u_j,\, r_b\rangle,\, \langle u_k,\, r_c\rangle\} \Rightarrow \{\langle u_l,\, r_d\rangle,\, \langle u_m,\, r_e\rangle\}$$

where u_i, u_j and u_k identify users and r_a, r_b and r_c are regions.

B. Observations

We formulate our association analysis [7] result as the two main observations:

1) **Locations of mobile phone users are associated.** Since 1868 user specified locations have been found (tuples) in the selected dataset, then we simply map each transaction into a vector of 1868 dimensions. The transactions throughout 59 days (1416 hours) are formed into a 1416×1868 matrix as the Figure 3a illustrated. Each pair of parallel segments indicates that these two user specified locations are associated; for instance when the user u_a stays at region r_i, the u_b always appears in the region r_j. For clear presentation, we try to capture one part (7 users within 48 hours) of the transaction matrix diagram in Figure 3b. It shows the locations of these users are associated. We found each user has two locations where are associated with each other's. Then we discover that most of these locations are in two types: (1) office, including MIT Media lab (cell tower 5119.408110 and 5119.403320) as well as their own working sites, and (2) their own residences. Finally we remark that the locations of users are associated by collective behaviors of crowds– i.e., going to work place in working time and staying at home in the time to rest.
2) **Association rules of CBP are supportless but confident.** We totally discovered 98853 association rules in selected transactions with support threshold of 1% and the confidence threshold of 10%. It is obvious that the most of rules are supportless, and only 79 rules are with the confidence larger than 10%. That means each of rules only exists in a limited number of transactions. However many rules are still confident. Inside of these 98853 rules, there are 82906 association rules with the confidence higher than 50%; and the confidence of 64621 rules is above 80%. Therefore, the association rules only exist in some part of users' locations, but it is confident to infer user's location from others' locations when the rule matches.

[1]The user #4, #8, #12, #23, #6, #69, #102, #37, #9, #26, #25, #27, #35, #73 and #53 are involved for empirical study.

From the observation 1 and 2 above, we believe that the *collective behavioral patterns* (**CBP**) and *association rules* of **CBP** are confident in inferring user's locations from others in the meanwhile. Driven by this objective, we have developed a location prediction model that considers the locations of all users, and predict their locations from the locations of other users for next 6 hours. We explain our algorithm in details next.

V. Location Prediction by Collective Behavioral Patterns

In this section, we present the mechanism to predict user's location by adopting *collective behavioral patterns*. Furthermore, we integrate our **CBP**-based schema with a Markov-based predictor [4] into a practical prediction schema regarding to both individual's mobility patterns and *collective behavioral patterns* associated with crowds.

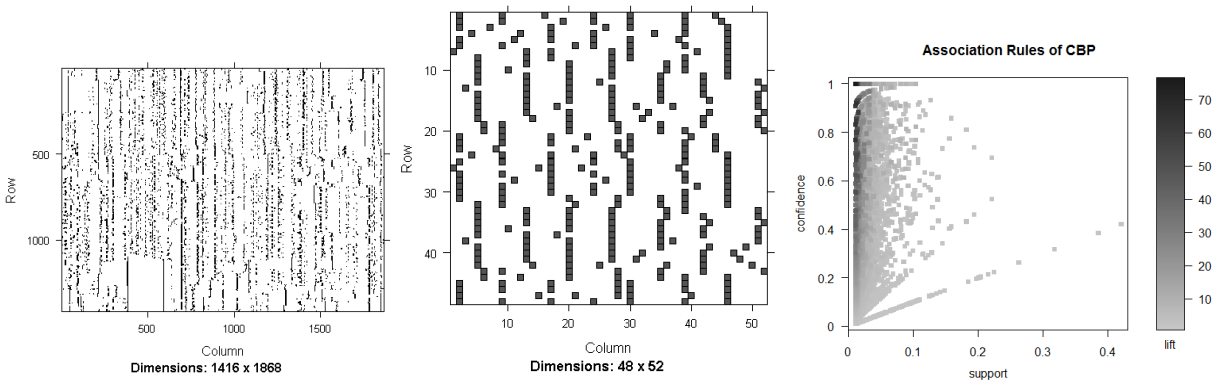
Although we have empirically observed the association rules of **CBP**, we still found that it is not appropriate to directly apply the association mining approaches and rule-based inference to location prediction, for many reasons. The main causes are: ***a)*** The association rules are not capable to infer a series of locations for one user in next 6 hours, since the **LHS** and **RHS** of a rule here reflect the locations of users in the meanwhile; ***b)*** Rules may have conflicts, e.g., $\{a, b\} \Rightarrow c$ and $\{a, b\} \Rightarrow d$, but the resolution way is not clear.

Hereby, our work proposes a Bayesian framework to predict user's locations to overcome above limits. Please make note that the inputs of **CBP**-based, Markov-based and the hybrid schemas are the GSM traces pre-processed as section III-A.

A. CBP-based Bayesian Predictor

In this section, we attempt at presenting the design of Bayesian schema to predict one user's future locations from the current locations of other users.

The predictor has two main components– candidate set generator and the probabilistic inference engine. (1) The candidate set generator yields a set of possible regions nearby user's *current location* first. (2) Then our predictor uses probabilistic inference engine to pick up the most



(a) Matrix Form of Transaction (b) Partial Enlargement (c) Association Rules of CBPs

Figure 3: Empirical Study

probable location from the candidate set as the prediction result for the next hour. (3) Since we assume the user would move to the location of prediction result in time, we assign user's current location with the prediction result, finally repeat step (1), (2) and (3) five more times for 6 hour prediction. Algorithm 1 addresses the skeleton of this predictor. The result of algorithm is an ordered list with 6 locations representing the locations for next 6 hours.

Algorithm 1 Skeleton of the Bayesian Predictor

$slot \leftarrow 0$
$current\_location \leftarrow$*user's current location*
$next\_location \leftarrow$**NULL**
$C \leftarrow \emptyset$
/*initiate the candidate set with an empty set*/
repeat
  $C \leftarrow gen(current\_location)$
  /*generate candidate set from current location*/
  $next\_location \leftarrow infer(C, slot)$
  /*forecast user's location by given time slot*/
  $list.add(next\_location)$
  /*add to the result list*/
  $current\_location \leftarrow next\_location$
  /*user is assumed to move to the next_location*/
  $slot \leftarrow slot + 1$ /*move to next hour*/
until slot=6
/*finish the prediction for six hour*/
return $list$

1) Candidate Set Generation: The Bayesian predictor picks up the most probable location from a candidate set of regions as the forecast result. Therefore, the generation of candidate set should have user's current spatial/temporal situation concerned. According to the definition 1 in section III, we have pre-processed data and extracted the previous trajectories of all users. Then we recover the topology of regions based on these trajectories. First, we collect the all neighbors of user's current locations in topology. Besides, user may still stay in the current location for future. Therefore the candidate set indeed is the collection of user's current location and its neighbors.

2) Probabilistic Inference Engine: Given the current locations of other users, i.e., the tuples $\langle u_1, r_1\rangle \dots \langle u_n, r_m\rangle$, the corresponding set of candidate locations C and the τ^{th} hour in future, this model is enabled to find out the most probable location:

$$\textbf{select } r \in C \quad \textbf{max } P(\langle u, r\rangle | \langle u_1, r_1\rangle \dots \langle u_n, r_m\rangle, \tau) \quad (1)$$

where u and r are the target user and the location (region) for prediction. $P(\langle u, r\rangle | \langle u_1, r_1\rangle \dots \langle u_n, r_m\rangle, \tau)$ identifies the conditional probability of user u staying at region r in future τ with the current locations of other users–i.e., $\langle u_1, r_1\rangle \dots \langle u_n, r_m\rangle$ given.

$$P_{bayes}(\langle u, r\rangle | \langle u_1, r_1\rangle .. \langle u_i, r_j\rangle .. \langle u_n, r_m\rangle, \tau) = \frac{P(\langle u, r\rangle) \times \prod_{i=1}^{n} \prod_{j=1}^{m} (P(\langle u_i, r_j\rangle, \tau | \langle u, r\rangle))}{P(\langle u_1, r_1\rangle .. \langle u_i, r_j\rangle .. \langle u_n, r_m\rangle)} \quad (2)$$

Suppose the locations between each of user pairs are weakly dependent, hereby it is reasonable to formulate the conditional probability in Eq. 1 with Naive Bayes modeling as Eq. 2.

$$P(\langle u, r\rangle) = \frac{\#\textit{records of } \langle u, r\rangle}{\#\textit{all records of user } u} \quad (3)$$

identifies the probability user r stay in region r.

$$P(\langle u_i, r_j\rangle, \tau | \langle u, r\rangle) = \frac{\#\textit{records of } \langle u_i, r_j\rangle \textit{ in the } \tau^{th} \textit{ hour before } \langle u, r\rangle}{\#\textit{ records of } \langle u, r\rangle} \quad (4)$$

is the probability that user u_i staying at r_j in the τ^{th} hour before user u staying in region r. Both probabilities are formulated in Eq. 3 and 4 with simple accumulation and

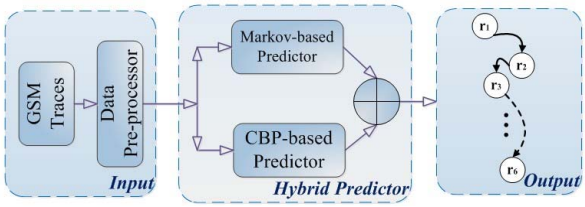


Figure 4: Architecture of Hybrid Predictor

statistics. If the acquired record don't exist–i.e.,$\#records = 0$, our predictor assigns $\epsilon = 0.000001$ as the value of these two probabilities in calculation. Furthermore, we simplify the Bayesian probability calculation as to estimate likelihood formulated in Eq. 5 which is proportional to the Bayesian probability.

$$\begin{aligned} &P_{bayes}(\langle u,r\rangle|\langle u_1,r_1\rangle..\langle u_i,r_j\rangle..\langle u_n,r_m\rangle,\tau) \\ &\propto \frac{P(\langle u,r\rangle)\times\prod_{i=1}^{n}\prod_{j=1}^{m}(P(\langle u_i,r_j\rangle,\tau|\langle u,r\rangle))}{\prod_{i=1}^{n}\prod_{j=1}^{m}P(\langle u_i,r_j\rangle)} \\ &= Likelihood(\langle u,r\rangle|\langle u_1,r_1\rangle..\langle u_i,r_j\rangle..\langle u_n,r_m\rangle,\tau) \end{aligned} \tag{5}$$

Finally the normalized likelihood shown in Equation 6 is employed as the approximation of conditional probability $P(\langle u,r\rangle|\langle u_1,r_1\rangle..\langle u_n,r_m\rangle,\tau)$. As supplementary, for each user u, we consider the region r that maximizes the normalized likelihood as the solution of Equation 1.

$$\begin{aligned} &P(\langle u,r\rangle|\langle u_1,r_1\rangle..\langle u_n,r_m\rangle,\tau)\approx \\ &\frac{Likelihood(\langle u,r\rangle|\langle u_1,r_1\rangle..\langle u_n,r_m\rangle,\tau)}{\sum_{r_k\in C}Likelihood(\langle u,r_k\rangle|\langle u_1,r_1\rangle..\langle u_n,r_m\rangle,\tau)} \end{aligned} \tag{6}$$

B. Integration with Markov-based Predictor

The **CBP**-based predictor is not designed to handle the main factor of human mobility–i.e., the individual's mobility pattern. Therefore our work integrates the **CBP**-based schema with a Markov-based predictor regarding to the individual's mobility patterns. The architecture of the hybrid predictor is illustrated in Figure 4. In this section, we mainly address the design issues of the markov-based predictor and the fusion process to combine the results of two predictors.

1) Markov-based Predictor: The location/trajectory predictors based on Markov chain model treat user mobility as sequences of locations, and take the individual's mobility patterns including the probability of transition between locations and the order of transitions into account. The implementation of our Markov-based location prediction schema is derived from [16] the state of the art of VMM-based sequence predictor.

For each user, we train a 6th-order markov chain to learn its mobility, and make the prediction by Partial Match (PPM) mechanism. This Markov-based predictor relies on the candidate set generation in section V-A1 also, and selects the most probable region from the candidate set C as the prediction result. The evaluation result section VI will show our Markov-based predictor delivers a sound prediction power, and it is a high quality baseline for performance comparison.

2) Prediction Result Fusion: The fusion process will be activated, when the prediction results of **CBP**-based and Markov-based schemas are different. Here we resolve the conflict through re-estimating the probability of each candidate region by using Evidence Theory [17]. The re-estimation is based on the joint mass (combination) calculation from the probability distributions of candidate regions given by above two predictors.

Thus, we consider the fusion process as the combination of evidences from both predictors based on Dempster's rule. In the framework of DS-Theory, mass functions are needed to give every possible set of results a degree of belief–i.e., $2^C \mapsto [0,1]$, where C is the candidate set of regions and 2^C is the power set of C. The masses for **CBP**-based predictor m_c and the masses for Markov-based schema m_m are formulated in Eq 7. $P_{CBP}(r)$ identifies the probability of region r for **CBP**-based prediction, and it is the same to P_{markov} for Markov-based predictor. Since the prediction result in our framework is a single region, so the masses of empty set and the set of multiple regions must be zero.

$$\begin{aligned} m_c(A) &= \begin{cases} 0 & |A|\neq 1 \\ P_{CBP}(r) & A=\{r\} \end{cases} \quad A\in 2^C \\ m_m(A) &= \begin{cases} 0 & |A|\neq 1 \\ P_{markov}(r) & A=\{r\} \end{cases} \quad A\in 2^C \end{aligned} \tag{7}$$

The fused probability from two predictors is formulated as the evidence combination as in Eq 8, where A *and* $B \subseteq 2^C$. According to the definition of masses in Eq 7, we can see the fused probability is actually proportional to the joint probability of **CBP** and Markov.

$$\begin{aligned} P_{fused}(r) &= (m_c(\{r\})\oplus m_m(\{r\})) \\ &= \frac{\sum_{A\cap B=\{r\}} m_c(A)m_m(B)}{1-\sum_{A\cap B\neq\emptyset} m_c(A)m_m(B)} \end{aligned} \tag{8}$$

Therefore, we can simply view the fusion result as the region r which can maximize the joint probability from both **CBP** and Markov.

Thus far, we have introduces the design and the implementation of three predictors – **CBP**-based, Markov-based and the hybrid schemas. The evaluation result presented in next section will further show the forecasting power of **CBP**-based schema and proves the capability of *collective behavioral patterns* to augment existing prediction approaches. We would be glad if interested users could contribute the other design of prediction schema regarding to the *collective behavioral patterns*.

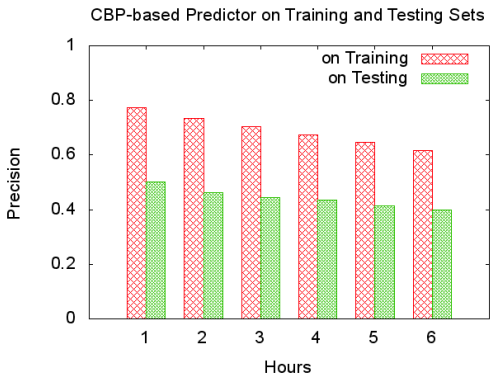


Figure 5: Aggregated Result: **CBP**-based Schema

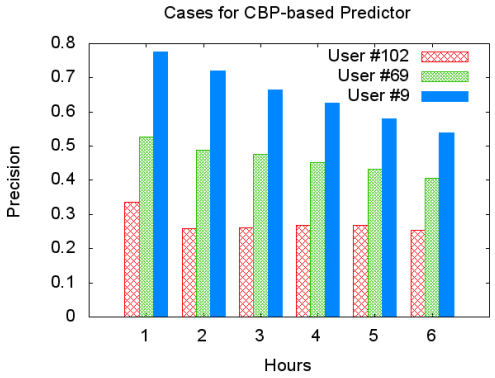


Figure 6: Cases of Three Users: **CBP**-based Schema

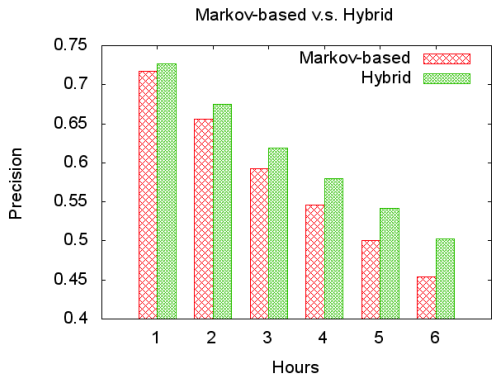


Figure 7: Aggregated Result: Markov-based v.s. Hybrid Schemas

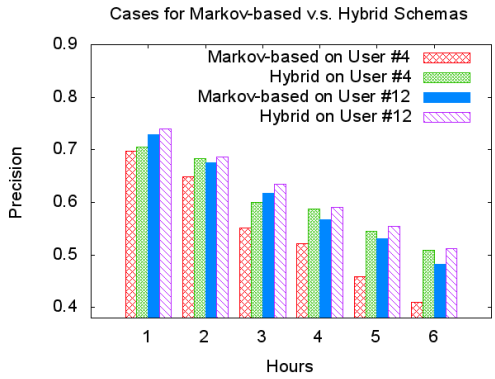


Figure 8: Cases of Two Users: Markov-based v.s. Hybrid Schemas

VI. Evaluation and Result

In this section, we show the evaluation results of our predictors on MIT dateset, and conduct a series of experiments to evaluate the prediction power of the **CBP**-based approach and its enhancement to Markov-based predictor.

In particular, we would like to answer: 1) whether **CBP**-based approach has ability to predict; and 2) in what degree the **CBP**-based schema can enhance the Markov-based predictor. Hereby, we start-up two sets of experiments, the first is the standalone evaluation of **CPB**-based schema, and another is performance comparison between Markov-based schema and the hybrid predictor. In this paper, we present the aggregated results of these experiments as well as the case study of few users.

Figure 5 illustrates the precision of **CBP**-based predictor running on the both training set and testing training. The precision on training set is decreasing from 77% to 62% for the prediction of 6 hours. It shows **CBP**-based schema conforms regularity of location prediction but is not overfitting. The precision on testing averagely declines from 50% to 40% in the prediction of 6 hours. Thus we believe **CBP**-based schema delivers sound quality of performance. Furthermore, the standalone performance of **CBP**-based schema for three users is addressed in Figure 6. These three cases are chosen as the worst (user #102), the average (user #69) and the best (user #9) cases of the standalone evaluation. It is reasonable to conclude that at least **CBP**-based schema is capable to achieve around 30% precision, however to the best case, the schema reaches more than 70% accuracy in the first hour prediction and decreases to the precision of 60% in the sixth hour prediction.

Figure 7 illustrates the aggregated performance comparison between Markov-based predictor (baseline) and hybrid schema. In average, our baseline–Markov-based schema delivers precision of 72% to 45% for the first to sixth hour prediction. It shows our hybrid schema outperforms Markov-based predictor averagely 1% in the first hour and around 5% in the sixth hour. The improvement of hybrid schema gradually raises for one to six hours. Besides the performance comparison between user #4 and user #12 is presented in Figure 8. The improvement on user #4 case goes from 1% to 10% for 6 hours; while the enhancement on user #12 cases remains no more than 3% for all hours.

The improvement of both cases are gradually increased by the hours for prediction, it is the same as what we have observed in aggregated comparison. We consider it is due to that **CBP**-based schema is able to correct the cumulative error in the sequence-liked prediction.

Reviewing above result and observation, it is reasonable to conclude that: 1) *collective behavioral patterns* can be used to predict mobile phone user's prediction, although they are not the main driver of human mobility, 2) our **CBP**-based schema is accurate and finally 3) *collective behavioral patterns* and our **CBP**-based schema can be adopted to augment the performance of Markov-based predictor or other predicting approaches.

VII. Conclusion

Location prediction has received substantive attention in recent years, yet it still remains to be addressed, particularly when it comes to mobile phone user mobility in urban environment. In this paper, we have proposed *collective behavioral patterns*, the association patterns among multiple users' locations in the meanwhile, provide a new perspective to leverage social aspects into user movement forecast. Moreover we design and implement a Bayesian prediction schema based on the *collective behavioral patterns*. Finally, **CBP**-based scheme also augments the performance of Markov-based predictor. Empirical studies over the continuous human mobility of 15 active users in two months demonstrate its effectiveness and accuracy.

However, **CBP**-based scheme could be further improved. We plan to design new prediction schemas adopting *collective behavioral patterns*. We will also validate the scalability and performance of **CBP**-based schema on millions of the realistic traces.

VIII. Acknowledgment

This work is supported by the EU FP7 Project SOCIETIES (No. 257493).

References

[1] B. Rao and L. Minakakis, "Evolution of mobile location-based services," *Communications of the ACM*, vol. 46, no. 12, pp. 61–65, 2003.

[2] A. Roy, S. Das, and A. Misra, "Exploiting information theory for adaptive mobility and resource management in future cellular networks," *Wireless Communications, IEEE*, vol. 11, pp. 59–65, 2004.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[4] F. Lassabe, P. Canalda, P. Chatonnay, F. Spies, N. Center, and D. Charlet, "Predictive mobility models based on kth markov models," in *IEEE Int. Conf. on pervasive services*, 2006, pp. 303–306.

[5] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *in Proc. of the 10th ACM intl. Sym. on MobiHoc*. ACM, 2009, pp. 299–308.

[6] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *in Proc. IEEE Intl. Conf. on Intelligent Transportation Systems, Madeira Island, Portugal*, 2010.

[7] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, pp. 55–86, 2007.

[8] A. S. P. N. Eaglea and D. Lazerc, "Inferring social network structure using mobile phone data," *in Proc. of the National Academy of Sciences (PNAS)*, vol. 106, pp. 15 274 – 15 278, 2009.

[9] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive wi-fi mobility data," in *in Proc. of INFOCOM 2004.* IEEE, 2004, pp. 1414–1424.

[10] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *in Proc. of the 17th ACM Conf. on KDD*, San Diego, CA, USA, 2011, pp. 1082–1090.

[11] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in *in Proc. of the 2nd Intl. Workshop on Multi-hop Ad Hoc Networks: From Theory To Reality*. Florence, Italy: ACM, 2006, pp. 31–38.

[12] C. Boldrini and A. Passarella, "Modelling spatial and temporal properties of human mobility driven by users' social relationships," *Comput. Commun.*, vol. 33, no. 9, pp. 1056–1074, 2010.

[13] K. P. A. H. W. J. Hsu, T. Spyropoulos, "Modeling time-variant user mobility in wireless mobile networks," in *in Proc. of the 27th IEEE Intl. Conf. on Computer Communications*, Alaska, USA, 2007, pp. 758–766.

[14] R. H. Katz, "CS-294-7: Handoff Strategies, University of UC Berkeley."

[15] Y. Liu, Z. Yang, X. Wang, and L. Jian, "Location, localization, and localizability," *Journal of Computer Sci. and Tech.*, vol. 25, pp. 274–297, 2010.

[16] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order markov models," *J. Artif. Intell. Res. (JAIR)*, vol. 22, pp. 385–421, 2004.

[17] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.