# Trajectory Enabled Service Support Platform for Mobile Users' Behavior Pattern Mining

Yanfeng Zhu, Yibo Zhang, Weixiong Shang, Jin Zhou, and Chun Ying
IBM China Research Lab
Email: zyfeng, zhangyib, shangwx, zhouzjin, yingchun@cn.ibm.com

***Abstract*—Existing operational support systems of GSM service providers focus on collecting and maintaining massive rough Cell-ID based location data, which cannot satisfy the requirement of new trajectory based services in identifying the behaviors of mobile users. In this paper, we introduce a trajectory enabled service support platform to convert the location data into limited meaningful mobile user's behavior patterns, which benefit the trajectory based services by simplifying the behavior detection. The core technologies of the platform are the pattern selection, which requires to cover the information included in the raw location data as more as possible, and the run-time mining algorithm, which requires less storage space. We propose a new concept, transient entropy, to identify the moving speed of users, and based on which we define and mine four types of behavior patterns: frequent locations, frequent trajectory, meaningful location, and moving mode. By analyzing the sojourn distribution, we find that the sojourn time in each location follows a Zipf distribution, based on which we present a run-time algorithm to mine the behavior patterns with less storage space. A realistic experiments is given to validate the proposed platform and algorithms.**

## I. Introduction

With the increasing development of mobile communication technology, mobile computing devices, especially cellphones, have become indispensable items of daily life. As indicated in the statistics from [20], in 2008, the number of cellphone users exceeds 2.5 billion, which is twice as many as the number of personal computer users worldwide. Meanwhile, GSM network, which is the main stream cellphone access network, has also been a ubiquitous sensor network, because the location of cellphone is equal to that of the user in most of time.

By sensing the location of a mobile user, mobile service providers and some third-part service providers can serve the mobile users better. In the last several years, many location based services [1][2][3] have been discussed and implemented. The core idea of these services is to utilize the current location of the mobile user to trigger a specific service operation, for example, sending a shopping advertisement to a mobile user moving around a store. The feedback from much realistic deployment shows that it is insufficient to deliver high quality services based on the current location only. For the same example, if the store locates on the user's way from home to office, the user will receive an advertisement everyday. It will cause a serious spam problem. To address this issue, mobile service providers move their attention to trajectory instead of pure location. A trajectory refers to a sequence of location records, which indicate the historical track of the mobile user. From the trajectory, the mobile service providers can mine several behavior features of the mobile user. Two typical trajectory based services are targeting campaign and mobile advertisement, which target to take precise operations according to the behavior features detected from the trajectory.

For a typical application scenario of trajectory based services, before taking service operations, the mobile service providers need to detect the behavior from the historical location data stored in database, for example, to check whether the mobile user goes to this location frequently. Therefore, extensive historical location data needs to be stored and analyzed. For a super mobile service provider with 100s million users, like China Mobile, it will be a big challenge in storage and run-time analysis.

Fortunately, the research on urban and economic geography [4][5][6] shows that the mobility-related phenomena and the meanings hidden behind the raw location data can be identified with limited temporal-spatial patterns, named Mobile User's Behavior Patterns (MUBP's) in this paper, which can be mined from the raw trajectory of a population of mobile users. Potential MUBP's include frequent locations (home, office, store in neighborhood, etc.), frequent trajectory (trajectories from home to office, from office to home, from home to supermarket, etc), meaningful locations (supermarket, cinema, and church frequently gone), moving mode (go with a private car or take public bus) and so on. The trajectory based services can query the service related information from the mined MUBP's directly without querying the massive historical location data. The main advantage of employing MUBP's is that:

- the storage space can be reduced significantly, from massive historical location data to limited MUBP's (the number is independent of time);
- it is easier to be used in the run-time applications because the operation over MUBP's brings less overhead in computation than that over massive historical location data.

The main challenges for employing MUBP's to serve trajectory based services are how to select a MUBP set to cover the information hidden behind the raw location data as much as possible and how to develop a time-space efficient pattern mining algorithm.

In this paper, we present a Trajectory Enabled Service Support Platform (TESSP), which is responsible for extracting MUBP's from the raw location data collected and maintained

by operational support systems (OSS) at run-time. Currently, we focus on four types of MUBP's: frequent location, frequent trajectory, meaningful location, and moving mode. Different from existing trajectory related pattern mining, we integrate the moving speed information in the MUBP mining by introducing a new concept – transient entropy. The moving speed based MUBP provides a new view to understand the behavior features of the mobile user from both time and space. By exploring the distribution feature of the location data, we propose a run-time algorithm for mining the defined MUBP's with less storage space. A realistic experiment based on 20 users' 3 month mobile location data is given to demonstrate the effect of the proposed TESSP.

The rest of the paper is organized as follows. Section II gives a brief overview to the proposed TESSP and the main challenges for the realistic implementation. In Section III, we propose an analytical model to investigate the features of the MUBP's and also the approaches for mining the MUBP's. The experiment is introduced in Section IV for validation. Related work is summarized in Section V. Finally, Section VI concludes the paper.

## II. Overview

In this section, we first introduce the TESSP, which is designed to extract MUBP's from the massive and rough location data. After that, we summarize the challenges for the realistic implementation,

### A. Trajectory Enabled Service Support Platform

The basic design target of the TESSP is to employ limited patterns to identify the essential behaviors hidden behind the massive and rough location data. The application services built upon the TESSP can quickly obtain the service related MUBP's by querying the pattern database directly. The architecture of the TESSP is shown in Fig.1.

The input of the TESSP is the mobile users' location data collected by OSS. There are two approaches for collecting the location data: one is to directly analyze the Cell-ID/LAC items in call/message records [17][18], and the other one is to employ client-side software to collect the cell handover logs [13][23][24][28][29][30]. The advantage of the former is that it is transparent for users, and do not bring any storage overhead to existing systems. However, no one always takes a call or sends a message, and thus this method cannot get complete trajectory information of users. In contrast, the latter involves the measurement in user side and requires additional storage for location information, but it is the best method to collect the complete location data. In this paper, we employ the latter method with client-side software developed in Java.

The output of the TESSP is the MUBP's stored in the mobility pattern database. Currently, we focus on four types of patterns: frequent location, frequent trajectory, meaningful location, and moving mode, which are considered to be the most important factors for trajectory based services.

Based on the available input and the targeted output, the TESSP employs 3 computational modules and 2 database storage modules to achieve the mining of MUBP's: data pre-processing module, data mining module, pattern template module, buffer database, and mobility pattern database. The basic functions are listed as follows:

- **Data pre-processing module:** is responsible for cleaning the raw location data, generating the location data and extracting the transient entropy which corresponds to the moving speed. After that, we obtain a triplet $\{time, location, transiententropy\}$ sequence to describe the state of mobile users in time and space.
- **Data mining module:** is responsible for mining the defined 4 types of patterns with the triplet based data.
- **Pattern template module:** is an interface for adding new pattern definitions to the platform. It is responsible for translating the requirement into triplet oriented rules. Due to the limited space, the detailed operational approaches are out of the scope of this paper.
- **Buffer database:** is employed to store the temporal triplet sequences and mediate statistic data, which will be used to mine the MUBP's.
- **Mobility pattern database:** is used to store the obtained MUBP's. The services can query the behavior patterns directly from this database.

The core technologies of the TESSP are conceived in the design of pre-processing module and the data mining module, which are also the main content of this paper.

### B. Challenges

The core idea of the TESSP is to convert the *massive and rough* location data into "limited" meaningful MUBP's at *run-time*. There are triple meanings / challenges in this process:

- the location data collected is rough, and thus it is challenging to utilize these low-precision location data to extract meaningful MUBP's;
- the location data collected by OSS is massive (usually hundreds records per day for one user), and thus it is challenging to select limited MUBP's to cover the information included in the raw location data as more as possible;
- the MUBP's are usually history related, and thus it is challenging to make the converting at run-time with as small storage space as possible.

The first challenge of TESSP results from the low-precision of data source. Different from GPS positioning, the location data in GSM network is a rough cell identity, Cell-ID, which just indexes a region covered by a base station. The basic principle of GSM positioning technology is that:

1) following GSM specifications, all base stations periodically broadcast cell information, which includes the Cell-ID;
2) cellphones utilize client-side softwares to collect the Cell-ID's of neighbor base stations and also measure the corresponding signal strength;
3) the cell with the maximum signal strength is considered to be the serving cell, and the user is assumed to

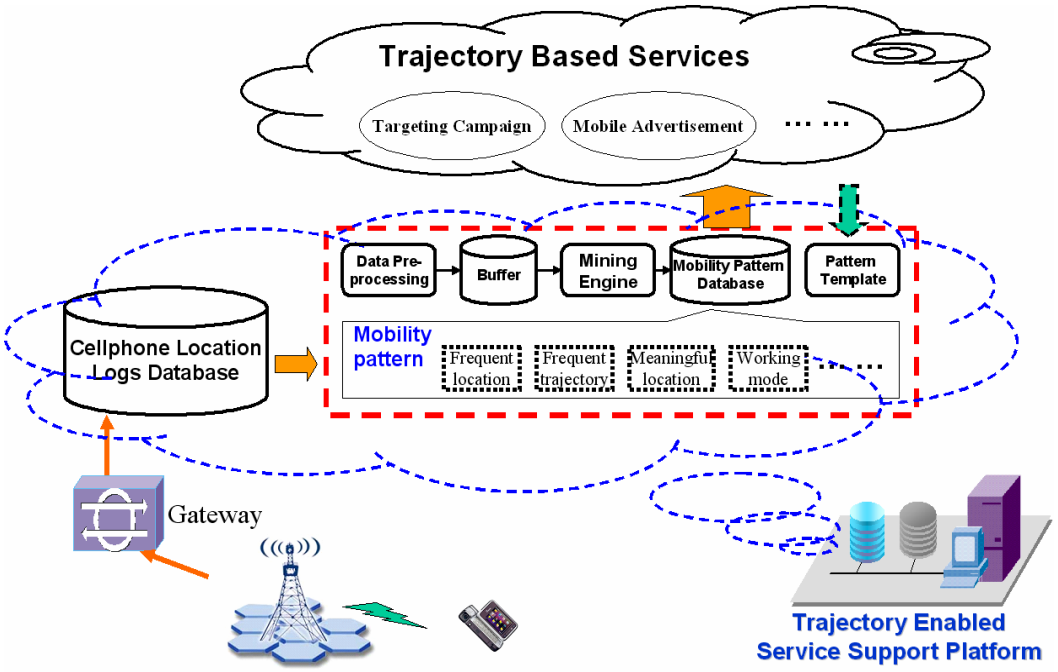



Fig. 1. Architecture of Trajectory Enabled Service Support Platform

TABLE I
User's Log Example

| Handover Time | Cell-ID | LAC |
|---|---|---|
| 2008-09-07 10:39:17 | 33665 | 4573 |
| 2008-09-07 10:41:31 | 33664 | 4573 |
| 2008-09-07 10:41:56 | 33665 | 4573 |
| 2008-09-07 10:42:11 | 58332 | 4573 |
| 2008-09-07 10:44:34 | 58331 | 4573 |

be staying in the coverage region of that cell. For convenience of description, we use the Cell-ID to denote the coverage of the cell.

In GSM network, client softwares [28][29][30] can track the handover points between neighboring BS's and record them in users' logs. Table.I shows an example of cell handover log. Herein, LAC (location-area-code) is the identity of a location manager, who maintains the location information of users within a location area. The location area consists of several cells, and the number depends on the user density and the realistic physical environment.

It is observed that Cell-ID positioning just output a cell centralized rough region. The precision varies with environment, ranges from 300 meters (urban) to several kilometers (suburban). Moreover, because the signal strength is time-varied with the randomness of wireless transmissions [31], the serving cell usually handovers among 2∼4 neighbor cells. As a result, the cell handover happens frequently even the user keeps static. Therefore, it is challenging to identify the behavior of mobile users. To address this issue, we extract the feature of moving speed to identify the details in a cell.

The second challenge comes from the selection of MUBP's. The location data collected is a sequence of cell sojourn record shown as Table.I. Intuitively, the frequent locations, usually the home, office, park in neighborhood etc, that appear frequently in the record are very important for mobile users. Besides that, given time and location, we can identify the moving mode of a mobile user, which is much useful to identify the meanings of a location for the mobile user. A user, who usually moves quickly with a private car[1], sojourning in a cell for 30 minutes implies that the user is taking some meaningful activities in the cell with high probability, this is usually named meaningful location. In addition, the frequent trajectories among meaningful locations are also of importance. To the best of our knowledge in telecommunication services, the four types of MUBP's summarized above have covered most of information in time and space, and they are enough to support most of existing trajectory enabled services.

Finally, the MUBP's are history related. the MUBP's are always hidden behind a long time statistic, for example, a park that a user goes every month. Unfortunately, the location data is massive, and thus it will cost a huge storage space for the history data. It is challenging to balance the requirement in statistic and storage space. To address this issue, all algorithms proposed in this paper are based on the run-time technology with less storage space.

## III. Model Analysis and Pattern Extraction

In this section, we present an analytical model to investigate the essential features hidden behind the raw location data.

[1]It costs approximately 1 minute in going through one cell by driving a car at speed 60kmph.

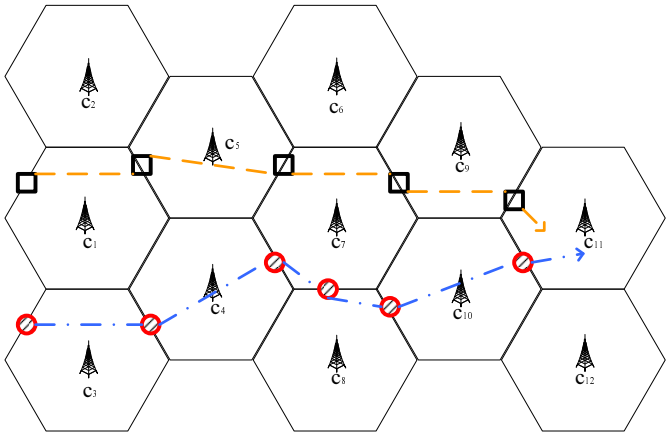



Fig. 2. Illustration of Typical Trajectory in Cellular Network

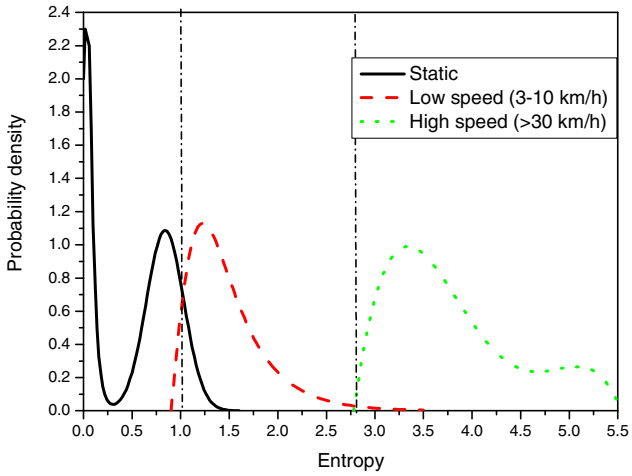



Fig. 3. Experimental Results for the Mapping Between Transient Entropy and Mobile Speed

Based on the analytical model, we provide the approach to extract four types of behavior patterns.

As mentioned above, Cell-ID positioning provides a rough cell based region information only instead of precise location information $(x, y)$ in GPS system. As shown in Fig.2, two trajectory examples are recorded as

$$C_1 \rightarrow C_5 \rightarrow C_7 \rightarrow C_9 \rightarrow C_{11}$$

$$C_3 \rightarrow C_4 \rightarrow C_7 \rightarrow C_8 \rightarrow C_{10} \rightarrow C_{11}$$

where $C = \{C_i, i = 1, 2, ..., N\}$ denotes the Cell-ID of each cell. Although the internal trajectory information within a cell is missed, the sojourn time in each cell, which is got from the difference between consecutive handover time points, can be employed to describe the behaviors in the cell. In addition, the handover time point is also very meaningful for understanding the behaviors of the user. Passing a supermarket at 9:00 am and that at 18:00 pm usually have different meanings for the user. Therefore, the handover time points should be retained in constructing the trajectory. Correspondingly, the user trajectory got from the GSM network consists of a sequence of cell sojourn records, shown as $< \{r_0, t_0^{in}, t_0^s\}, \{r_1, t_1^{in}, t_1^s\}, ..., \{r_i, t_i^{in}, t_i^s\}, ... >$. Herein, $r_i \in C$ is the Cell-ID in $i$th sojourn record, $t_i^{in}$ is the handover time point that the user handovers from other cells to $r_i$, and $t_i^s$ is the sojourn time in $r_i$, i.e., $t_i^s = t_{i+1}^{in} - t_i^{in}$. The trajectory can also described as

$$\{r_0, t_0^{in}\} \xrightarrow{t_0^s} \{r_1, t_1^{in}\} \xrightarrow{t_1^s} .....\{r_i, t_i^{in}\} \xrightarrow{t_i^s} .....$$

### A. Triplet for User's State

As mentioned above, the moving speed, which is much useful for understanding the behaviors of users, should be extracted from the raw location data before mining the behavior patterns. However, it is hard to compute the precise speed with the cell sojourn records due to several reasons:

- the cell radius varies with the deployment environment, $0.3 \sim 1$ km in urban areas and $1 \sim 10$ km in suburban areas;
- due to the randomness of wireless signal transmissions, the ping-pong effect[2][27] happens frequently in the overlapping region of neighbor cells, which results in many randomly handovers even the user is static.

Alternatively, we consider a fuzzy approach to approximate the moving speed. An intuitive solution is to use the number of cells passed in a unit time. However, the ping-pang effect will cause much error. Instead, we propose a new concept, transient entropy, to approximate the mobile speed. The transient entropy is defined as follows:

$$H(t, \Delta T) = -\sum_{i=1}^{N} P_{t,\Delta T}(C_i) \log_2 P_{t,\Delta T}(C_i) \quad (1)$$

Herein,

$$P_{t,\Delta T}(C_i) = \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} I(C_i, r_\tau) \mathrm{d}\tau \quad (2)$$

where $r_\tau$ denotes the Cell-ID of the cell that the user stays at time $\tau$ and

$$I(C_i, r_\tau) = \begin{cases} 1, & C_i = r_\tau, \\ 0, & \text{others.} \end{cases}$$

Because the transient entropy takes the sojourn time into account, it is much better than cell number counting.

Fig.3 shows the experimental results about the mapping relationship between transient entropy and realistic mobile speed. The data comes from our realistic experiment. It is observed that the transient entropy in static state has two peaks: one is at 0 and the other is near 0.8, which is due to the ping-pang effect. By setting some thresholds, the moving speed can be clarified clearly with the transient entropy.

Correspondingly, given $\Delta T$, the state of the user in time $t$ can be described by a triplet $\{t, r_t, H(t, \Delta T)\}$. To give an

[2]In the overlap of two neighboring cells, the mobile user handovers frequently due to the randomness of wireless signal transmissions even the the mobile user does not move.

intuitive expression for the trajectory of a user, we plot the variation of triplet in 24 hour. Fig.4-(a) shows the result in a 3-D visualization mode. Herein, to better show the figure, we convert the real Cell-IDs into a sequence of fake Cell-IDs, which range from 1 to N. Note that the value of Cell-ID does not have any meanings in distance, just for indexing. Fig.4-(b,c,d) show the projections in different planes, respectively.

- Fig.4-(b) shows the relationship between location (Cell-ID) and time. It is observed that the user has an obvious office worker behavior mode: stay a long time at two places (home and office); move at morning (may be from home to office), noon (may be out for lunch), and afternoon (may be from office to home). Note that the frequent handovers at 9:00-13:00 and 14:00-19:00 is due to the ping-pang effect.
- Fig.4-(c) shows the relationship between transient entropy and time. It is observed that there are two entropy peaks at 8:00-9:00 and 19:00-20:00, and the peak values indicate that the user was moving in high speed ($> 30$ km/h), which was confirmed with the user that he was on the way to office and home.
- Fig.4-(d) shows the relationship between transient entropy and location (Cell-ID), which indicates the speed that the user passes each cell. It is observed that the user frequently appears in a limited number of locations with low speed, and pass other locations with high speed. This figure is much useful to identify the frequent location and meaningful locations.

### *B. Pattern Extraction*

Here, we explore the methods for identifying various MUBP's.

*1) Frequent Location:* The identification of frequent location is based on the sojourn time ratio, which is defined as the ratio between the cumulative sojourn time in this cell and the total monitor time. The sojourn time ratio for $C_i$, $R(C_i)$, is given by

$$R(C_i) = \frac{1}{T}\int_0^T I(C_i, r_\tau)\mathrm{d}\tau, \tag{3}$$

where $T$ is the total monitor time. As defined in the last section, the locations with the top $N_l$ longest cumulative sojourn time are considered to be the frequent location.

*2) Meaningful Location:* The identification to meaningful locations is based on two rules:

- the sojourn time ratio is larger than a threshold $\theta_M$;
- the average transient entropy in this cell is less than a threshold $\phi_M$

The average transient entropy for $C_i$, $\bar{H}(C_i, \Delta T)$ is given by

$$\bar{H}(C_i, \Delta T) = \frac{1}{TR(C_i)}\int_0^T H(\tau, \Delta T) I(C_i, r_\tau)\mathrm{d}\tau \tag{4}$$

Then, the cell, which satisfies the following conditions, can be identified as meaningful location.

$$\begin{cases} R(C_i) \geq \theta_M \\ \bar{H}(C_i, \Delta T) \geq \phi_M \end{cases} \tag{5}$$

*3) Frequent Trajectory:* For the mining of frequent trajectory, we focus on the trajectories among meaningful locations only. From the viewpoint of application, these trajectories are more sensitive for service providers. From the viewpoint of mining technology itself, the meaningful locations are frequent, and thus the trajectories among them are frequent with high probability. Among these trajectories, we select the trajectories with the top $N_t$ frequency as the frequent trajectory.

Traditional trajectory mining focus on matching the trajectory according to the order of the location sequence[21][22]. Differently, we ignore the order, which can reduce the computation complexity significantly. We employ the meaningful locations to identify the start and end points first. It is well known that the physical positions of cells are fixed. Taking Fig.2 for example, for the trajectories from $C_1$ to $C_{11}$, if two trajectories are the same, the order of the location sequences should be the same with high probability. Therefore, we ignores the order in mining process, and take a simple similarity matching for instead. For two trajectories ($\chi_1$ and $\chi_2$) with the same start and end points, we consider them as the same if and only if

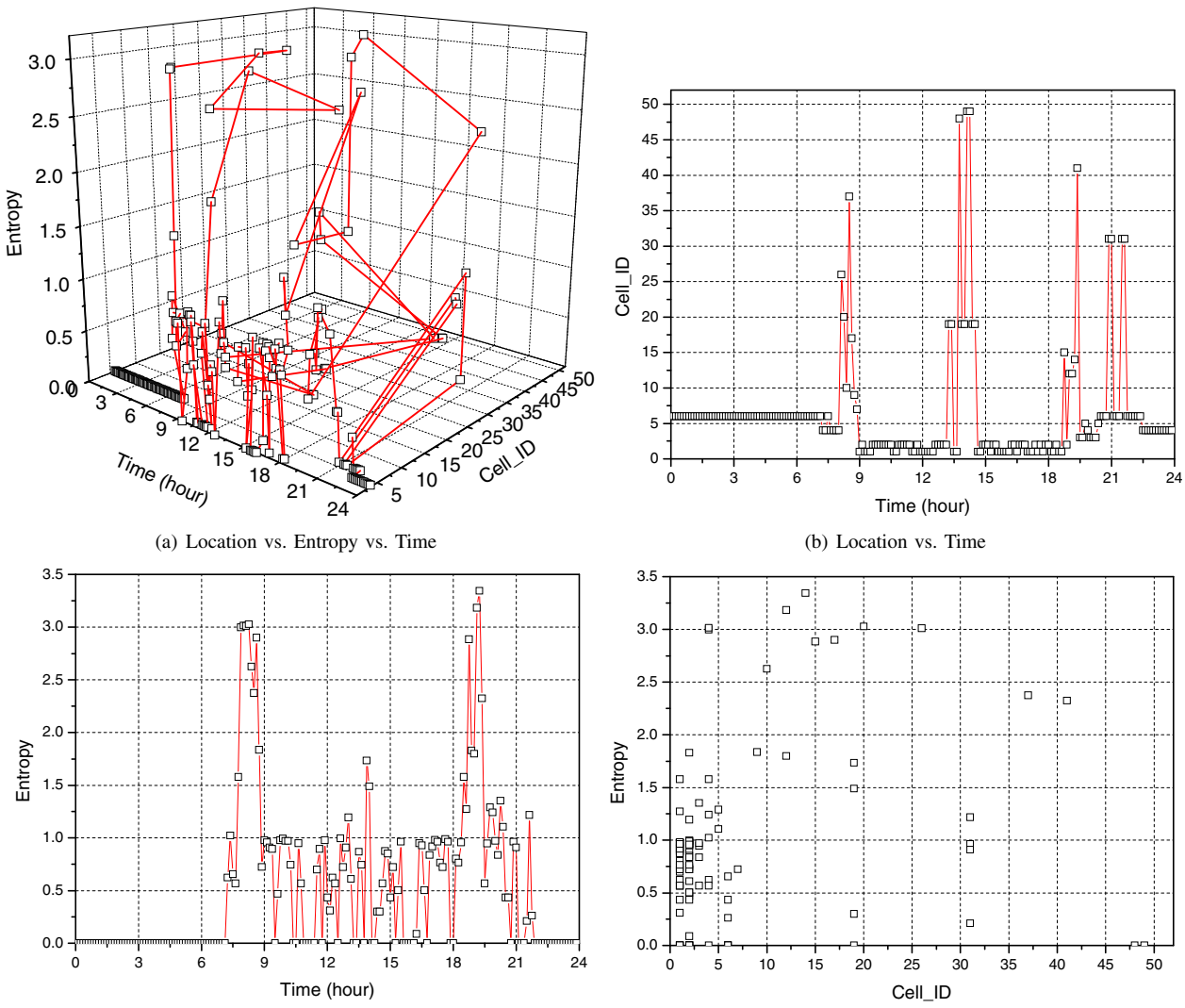
$$\frac{N(\chi_1, \chi_2)}{\min\{N(\chi_1), N(\chi_2)\}} \geq \gamma \tag{6}$$

where $N(\chi_1, \chi_2)$ denotes the number of cells included in both $\chi_1$ and $\chi_2$ and $N(\chi_i)$ denotes the number of cells in $\chi_i$. Especially, $\gamma = 1$ refers to that only the trajectories with the same cell sojourn records are considered to be the same. In the experiment of this paper, we set $\gamma = 0.8$ to alleviate the impact of overlapping issue[3].

*4) Moving Mode:* The moving mode in this paper refers to the transportation tools: by private car or not. The essential difference between the users with private car and those without can be observed in the average moving speed and the distribution of moving speed in moving state.

Intuitively, it sounds enough to employ the average moving speed, which is quantified by the transient entropy, to distinguish the moving mode. However, the experimental result shows that the error is not ignorable. As mentioned in the definition of transient entropy, the value of transient entropy is related to both the physical moving speed and the cell density. Therefore, the average transient entropy varies with both the transportation tool and the moving region: the users with private cars and living in suburban regions, where the cell density is lower, usually has smaller average transient entropy than those without private cars but living in urban regions.

[3]As mentioned above, a physical location is usually covered by multiple cells, which results in that the cell sojourn records are different even passing the same physical location.




(a) Location vs. Entropy vs. Time

(b) Location vs. Time

(c) Transient Entropy vs. Time

(d) Transient Entropy vs. Location

Fig. 4. Relationship among Time, Location, and Transient Entropy

Alternatively, the distribution of moving speed, which corresponds to the distribution of the transient entropy, is also useful for the differentiation. The users without private cars take more time in low-speed movement (by walk or bicycle), and differently the users with private cars usually cost more time in high-speed movement. Therefore, the sloping of the probability density line of the transient entropy varies with the moving mode.

Herein, we concentrate on the moving mode in moving state, and thus we first remove the part that the users stay in static state by filtering the parts with $H(t, \Delta T) < 2.0$. Fig.5 shows an experimental result for the transient entropy distribution of two users: one with a private car and the other is not. We employ a linear fitting function to fit the probability density line of each user, the linear fitting function is

$$p_d = f_s H + f_c \tag{7}$$

where $H$ is the transient entropy, $p_d$ is the probability density, and $f_s$ is the slope of linear fitting line.

In summary, we can employ any one of the following rules to identify the moving mode of users (whether have a private car):

- the average transient entropy is larger than a threshold $\phi_a$ in moving state, where the moving state is defined as the state with the transient entropy being larger than a threshold $\phi_s$ (in this paper, we set $\phi_s = 2.0$);
- the slope of the probability density line of the transient entropy is smaller than a threshold $\kappa$.

In the following experiments, we find that employing a joint detection with both rules above can significantly reduce the classification error.

*5) Discussion:* The identification of meaningful locations and trajectories highly depends on the moving mode of users. It is straightforward that a low-speed movement of the user, who has a private car and always moves fast, usually means that he is taking some meaningful activities, for example, stop

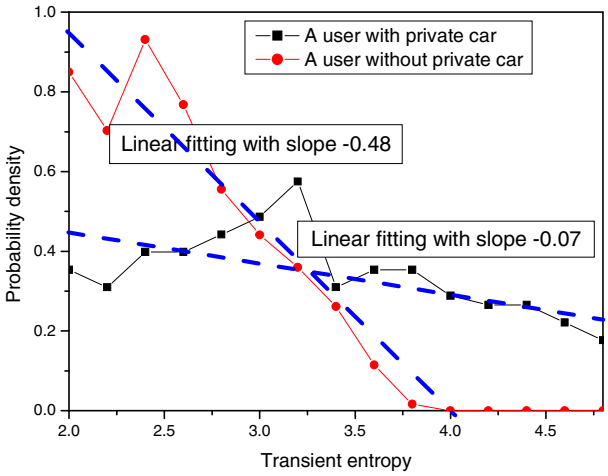



Fig. 5. Illumination of Transient Entropy Distribution with Different Moving Mode

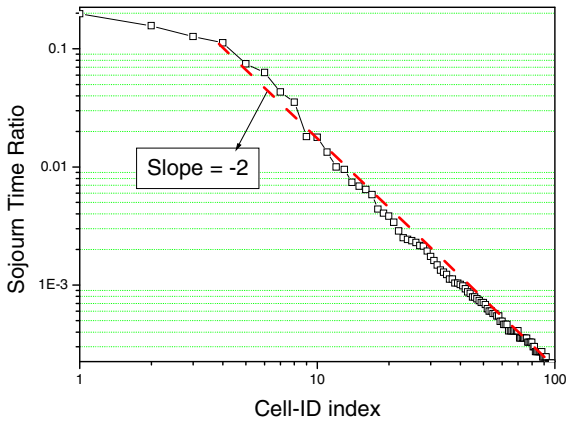



Fig. 6. Zipf Distribution of Sojourn Time Ratio

to enjoy a cup of coffee or enter a supermarket. Differently, for a user without a private car, the movement with the same speed usually means that he is still on the way home.

Therefore, the identification of moving mode is one of important factors to decide the setting of these thresholds used in the identification of meaningful locations and trajectories.

### C. Run-Time Implementation with Smaller Storage Space

The population of mobile users is very large, for example, more than 10,000,000 users in Beijing, China. In our experiment, we find that there are more than 300 cell handovers each day (in the data set of [13], there are more than 200 cell handovers each day). As shown in Table.I, one cell handover record costs at least 16 bytes, and then it will cost at least 48G bytes storage space each day. As a result, it is difficult to take an on-line analysis to mine some monthly behaviors, which require at least one month storage. To address this issue, we consider a run-time algorithm to leverage the storage requirement.

The basic principle of the run-time algorithm is to measure the local statistics of the sojourn time ratio and the transient entropy at a short interval and then update to the overall statistics with a smooth factor. Taking the sojourn time ratio as example, the run-time statistics are shown as

$$R(C_i) = \frac{T-\delta}{T} R(C_i) + \frac{\delta}{T} \bar{R}(C_i) \tag{8}$$

where $T$ is the targeted overall statistical period[4], $\delta$ is the short interval for updating at run-time, and $\bar{R}(C_i)$ is the local statistics of the sojourn time ratio in the latest short interval.

In the TESSP, the top $N_l$ locations, the top $N_t$ trajectories, meaningful locations and the moving mode (depends on the distribution of the transient entropy) are required to be mined based on the overall statistics. It is clear that a top $N_l$ location in a short interval $\delta$ does not means that it will be a real top $N_l$ location in the overall statistics. Therefore, more than $N_l$ cells are required to be maintained to calculate the overall top $N_l$ locations. A straightforward solution is to maintain the statistics of all cells, which can completely avoid the missing of the overall top $N_l$. However, there are 1000s∼10,000s cells in an urban region, it is cost-inefficient to maintain the statistics of all cells.

Fortunately, the statistics over our realistic experiment show that the distribution of the sojourn time ratio in each cell follows the Zipf law [32]:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)} \tag{9}$$

where $k$ is ordered index for the considered variable and $k = 1, 2, ..., N$. Zipf distribution can be detected with

$$\log f(k_1; s, N) - \log f(k_2; s, N) = -s(\log k_1 - \log k_2) \tag{10}$$

In Fig.6, we show that the sojourn time ratio distribution, which is the result averaged over all users. Herein, the x-axis is the Cell-ID index ordered by the sojourn time ratio, and y-axis is the value of sojourn time ratio. It is observed that the sojourn time ratio follows Zipf law with $s = 2$ starting from the 4th cell. The reason that the top 3 cells do not follow Zipf law can be explained with the truth that a physical position is usually covered by several cells.

Following the Zipf law, the sojourn time ratio of the $N_l$th is $\left(\frac{M_l}{N_l}\right)^2$ times of that of the $M_l$th cell. If we want to maintain the statistics to top $M_l$ cells in the short interval $\delta$ only, the overall top $N_l$ over $T$ can be retained as long as $M_l \geq \left(\frac{T}{\delta}\right)^{1/s} N_l$. The top $N_t$ trajectories can be calculated with the similar method as well. For the mining of meaningful locations, which requires to get all locations those sojourn time ratio being larger than $\theta_M$, the threshold is relaxed to $\frac{\delta}{T}\theta_M$ in each update interval $\delta$. In the following experiment, we set $T = 30$ days and $\delta = 3$ days, i.e., we just need to store 3 days data. Correspondingly, $\left(\frac{T}{\delta}\right)^{1/s} = 3.16$, and for mining top 10 frequent locations, only the statistics to top 32 cells are maintained. The overhead is acceptable.

[4]For the requirement of mining monthly behavior, it is at least one month.

## IV. Experiment Validation

Before diving into the detailed experimental analysis, we first introduce the experimental environment in detail. After that, we describe a solution for alleviating the impact of the missing data on the model. After that, The experimental result for each type of MUBP is given.

### A. Experimental Environment

Similar to the experiment in MIT [13], we select 20 IBM China research lab staffs for daily location monitoring. Each staff takes a Nokia 7650 smart phone pre-installed with a Cell-ID handover recording software developed with Java. The experimental environment is the urban region of Beijing in China. Among these staffs, 10 take private cars as the main transportation tool, and others walk or take public transportation tool–bus as the main transportation tool. The Cell-ID handover records are the same to that shown in Table.I. The experimental data covers 3 months with approximately 43,200 hours.

### B. Missing data

In our experiment, the Cell-ID records of each staff are uploaded to a server every morning, and a missing data checking program will be used to search and fix the missing data. The missing data usually results from entering signal blind region and power off. The principle of fixing the missing data is that:

- First, we check whether the behavior of the target user has obvious period by spectrum analysis over transient entropy [5]. Before taking the spectrum analysis, we separate the data in weekday and weekend because they intuitively have different behavior modes.
- If the power spectrum has obvious spectrum peak in day-period like that shown in Fig.7, it is considered to be that the target user has periodical behavior. We select the data from last several days falling in the same time duration as the missing data to fix it.
- If there is no peak in day-period, the miss data part is kept to be void.

In our experiment, the ratio of missing data and overall data is nearly $5\%$. Therefore, fixing the missing data is much useful to avoid the meaningful locations are misjudged due to the cut sojourn time ratio.

### C. Experimental Results

As analyzed in Section III, the thresholds for judging meaningful location and trajectory depend on the moving mode of users, and thus we first identify the moving mode. Based on the result of moving mode identification, we investigate the meaning locations and the frequent trajectories. Because the mining result of the frequent locations is straightforward, we do not take additional experiments to validate it.

[5]The spectrum analysis over Cell-ID is also workable, but the physical meaning is not clear.

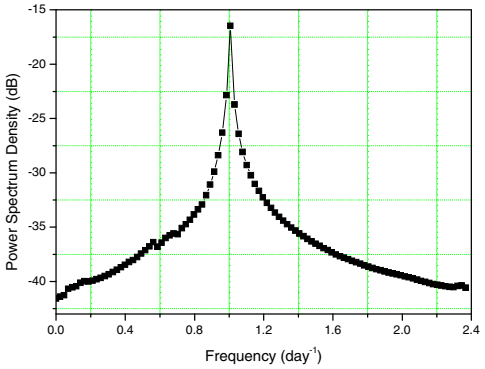



Fig. 7. Power Spectrum of A User's Transient Entropy

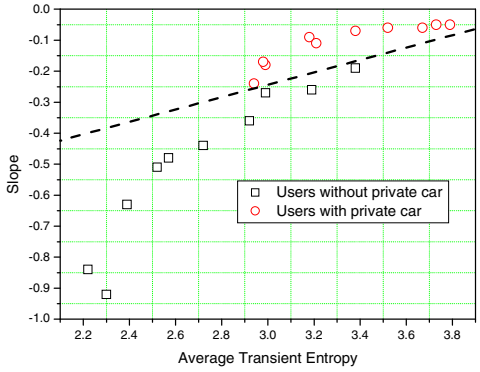



Fig. 8. Moving Mode Classification Based on Average Transient Entropy and Fitting Slope

*1) Moving Mode:* We first filter all records in static state and only make statistics to the samples with entropy being larger than 2. The slope $f_s$ of fitting equation (7) and average transient entropy $\bar{h}$ of each staff is plotted in Fig.8. It is observed that it is impossible to classify the moving mode completely with slope $f_s$ or average transient entropy only. This is mainly because staffs reside in different regions, which correspond to different cell deployment densities, and therefore a sole rule is not always workable. For instead, a linear function can be employed:

$$f_s = 0.21\bar{h} - 0.87 \tag{11}$$

*2) Meaningful Location:* As mentioned above, the judgement of meaningful location depends on the moving mode. The threshold for the average transient entropy $\phi_M$ is selected according to $\bar{h}$, which is used to identify the moving mode. In this paper, $\phi_M$ is set as

$$\phi_M = \bar{h} - 2. \tag{12}$$

To give an intuitive expression for the difference between frequent locations and meaningful locations, we first plot the results of one staff with $\bar{h} = 2.39$ in Fig.9. Correspondingly, $\phi_M = \bar{h} - 2 = 0.39$. Herein, $\theta_M = \frac{0.5}{24\times 7} = 0.003$, which corresponds to staying half hour per week (typically, it is the sojourn time for supermarket and gymnasium those gone weekly). It is observed that some locations with high transient

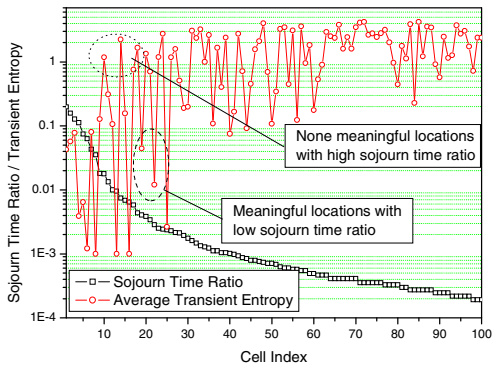



Fig. 9. Distribution of Sojourn Time Ratio and Average Transient Entropy

entropy are filtered by the proposed detection method. By confirming with the related staff, these locations fall on the way between home and office, and he just passes them by bus.

In Table.II, we summarize the experimental results of meaningful location detection of the staffs attending the experiment. Herein, the last 10 staffs have private car, and the first 10 do not have. The number of meaningful locations detected by our algorithm ($n_d$) is recorded in third column, the number of meaningful locations confirmed by the staffs ($n_c$) is listed in the fourth column, and the number of meaningful locations of the staffs missed by the detection ($n_m$) is listed in the last column. Especially, the total number of meaningful locations of the staff is $n_c + n_m$, and the accurate of the detection is $\frac{n_c}{n_d}$. Therefore, the metric, false-positive ($P_{fp}$), can be used to evaluate the performance of the proposed approach:

$$P_{fp} = \frac{n_c}{n_c + n_m} \times \left(1 - \frac{n_c}{n_d}\right) \tag{13}$$

From Table.II, the false-positive averaged over the staffs with private car is 0.0490, and that average over the staffs without private car is 0.0557. Intuitively, the behaviors of the staffs with private car are easier to be detected than those without private car, because their moving and stop states are easier to be distinguished.

The results show that most of staffs have some none meaningful locations with higher transient entropy. By leveraging $\theta_M$ and $\phi_M$, the detection of meaningful locations can exploit more sense hidden behind the raw location data. Especially, the result got with $\phi_M = \infty$ is equal to that got from frequent location mining. Contrastingly, the result based on $\theta = 0$ is that mining all locations that the user moves slowly.

*3) Frequent Trajectory:* We summarize the frequent trajectories of a staff in Fig.10, where solid lines shows the frequent trajectories and the thickness indicates the frequency. In addition, the dash line shows some other trajectories out of top $N_t$. It is observed that the topologies of the frequent trajectories follow star mode with multi-center. This feature implies that a optimization in storage can be employed in the mining process.

Currently, we need to maintain all trajectories among meaningful locations. For a user with 20 meaningful locations, we need to maintain 380 trajectory sets for the selection of frequent trajectory. With the star topology, only the trajectory sets between the centric meaningful locations and other meaningful locations are maintained. For a star with two-center topology shown in Fig.10, only 36 trajectory sets need to be stored. Due to the limited space, the detailed optimization solution will be discussed in our future work.

TABLE II
Summary of Meaningful Location Detection

| User | $\phi_M$ | Detected | Confirmed | Missed |
|---|---|---|---|---|
| 1 | 0.57 | 20 | 18 | 4 |
| 2 | 0.52 | 20 | 17 | 7 |
| 3 | 0.72 | 18 | 16 | 6 |
| 4 | 0.39 | 18 | 17 | 7 |
| 5 | 0.30 | 19 | 17 | 6 |
| 6 | 0.92 | 20 | 19 | 4 |
| 7 | 1.38 | 21 | 21 | 2 |
| 8 | 0.22 | 13 | 12 | 4 |
| 9 | 0.99 | 21 | 19 | 6 |
| 10 | 1.19 | 20 | 20 | 3 |
| 11 | 1.38 | 21 | 19 | 1 |
| 12 | 0.98 | 18 | 16 | 5 |
| 13 | 1.67 | 20 | 19 | 2 |
| 14 | 1.18 | 20 | 18 | 2 |
| 15 | 1.73 | 23 | 23 | 1 |
| 16 | 1.79 | 22 | 21 | 2 |
| 17 | 0.94 | 18 | 17 | 4 |
| 18 | 1.52 | 20 | 20 | 1 |
| 19 | 1.21 | 21 | 21 | 0 |
| 20 | 0.97 | 17 | 15 | 4 |

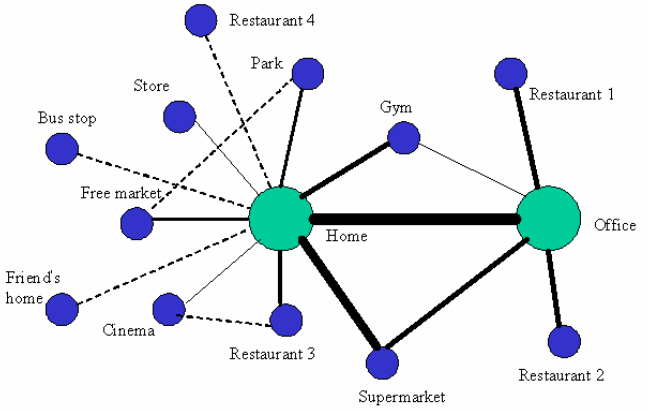



Fig. 10. Frequent Trajectory of A Staff

## V. Related Work

For the trajectory mining related work, most of them [8][9][10][11][12] focused on the location from GPS receivers. Because GPS receivers can theoretically provide precise location data, and thus these work paid more attention on the precise frequency sequence pattern mining / clustering technologies. However, mass deployment of GPS receivers is very expensive, and the transmission cost to the central system and the power consumption to mobile clients are proved to be onerous [16]. Moreover, GPS systems do not work well in

urban areas due to no line-of-sight path. Therefore, GPS receivers are not a suitable solution for mobile service providers. The location data from GSM, which is just rough Cell-ID data indicating a 100s-1000s meters range, is much different from GPS based location data. The algorithms developed in these work cannot be applied.

Based on the Cell-ID positioning, the authors of [13][15] take many experiments to identify frequent locations and meaningful locations. Because they did not take the moving speed into account, the judgement of frequent locations and meaningful locations depends on the sojourn time only but ignoring the moving speed based patterns which hidden behind the location data sequence. It is well known that a location that the considered user pass by walk and by car imply different meanings: the former implies that the considered user is a potential customer of a nearby supermarket with high probability, but the latter implies the probability is very low. To the best of our knowledge, this is the first paper to involve the moving speed in the pattern mining.

## VI. Conclusion

In this paper, we concentrate on the mining of mobile user's behavior patterns in GSM network. A platform is introduced to employ limited patterns to cover the information included in the massive and rough location data. Different from existing location based mining work, we present a transient entropy to identify the moving speed of users in complicated cell deployment environment. Then, the state of user in time and space is described with a triplet $\{$time, location, transient entropy$\}$. Based on the triplet, four types of patterns are defined. Utilizing the distribution feature of sojourn time in locations, we implemented the mining algorithm at run-time with less storage space. The main contribution of this paper is summarized as:

- presenting a platform to converting the massive and rough location data into limited mobile user's behavior patterns;
- employing the transient entropy to identify the moving speed of mobile users, with which the locations the user always passes with high speed can be filtered from the meaningful locations;
- developing a run-time mining algorithm to alleviate the requirement in storage space by utilizing the distribution feature of sojourn time;

## Acknowledgment

The authors would like to thank Professor Cecilia Mascolo for discussion and the contribution to refine this manuscript. Additionally, we thank Ping Pan and Ling Jin for their contribution in trajectory data collection.

## References

[1] Stefan Steiniger, Moritz Neun, and Alistair Edwardes, "Foundations of Location Based Services," , University of Zurich.

[2] GSM Association, "Permanent Reference Document SE.23: Location Based Services".

[3] Shu Wang, Jungwon Min, and Byung K. Yi, "Location Based Services for Mobiles: Technologies and Standards," IEEE International Conference on Communication (ICC) 2008, Beijing, China, 2008.

[4] Golledge R G, "Stimson R J., Spatial Behavior: A Geographic Perspective", New York: The Guilford Press, pp. 349 - 386, 1997.

[5] Szalai A, et.al., "The use of time: daily activities of urban and surburban population in twelve countries", The Hague: Mouton, 1966.

[6] Martin E.H., Lee-Gosselin, and Sean T. Doherty, "Measuring activity and action space/time: Are our methods keeping pace with evolving behaviour patterns? In Integrated Land-Use and Transportation Models: Behavioural Foundations," pp. 101-132, Oxford: Pergamon-Elsevier.

[7] J. Han, J. Lee, H. Gonzalez, and X. Li, "Mining Massive RFID, Trajectory, and Traffic Data Sets", SIGKDD 2008 tutorials, August 2008.

[8] F. Giannotti, M. Nanni, and D. Pedreschi, "Efficient mining of sequences with temporal annotations," In Proc. SIAM Conference on Data Mining, pp. 346C357, SIAM, 2006.

[9] F Giannotti, M Nanni, F Pinelli, D Pedreschi, "Trajectory Pattern Mining", SIGKDD 2007, August 2007.

[10] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory Clustering: A Partition-and-Group Framework," SIGMOD'07, pp. 593-604, 2007.

[11] H. Cao, N. Mamoulis, and D. W. Cheung, "Mining frequent spatio-temporal sequential patterns," ICDM'05, Nov. 2005.

[12] Y. Cai and R. Ng, "Indexing spatio-temporal trajectories with chebyshev polynomials," SIGMOD'04, pp. 599-610, 2004.

[13] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems", Personal and Ubiquitous Computing, vol. 10, no. 4, 255-268, 2006.

[14] K. Laasonen, M. Ranto, and H. toivonen, "Adaptive on-device location recognition," in *Proceedings of the Second International Conference on Pervasive Computing*, pp. 287-304, Springer, 2004.

[15] P Nurmi, J Koolwaaij, "Identifying Meaningful Locations", Mobile and Ubiquitous Systems-Workshops, 2006.

[16] Nico Deblauwe and Peter Ruppel, "Combining GPS and GSM Cell-ID positioning for Proactive Location-based Services", Mobile and Ubiquitous Systems-Workshops, 2007.

[17] Francesco Calabrese, *et.al.*, "Real-Time Urban Monitoring Using Cellular Phones a Case-Study in Rome," techinical paper, senseable.mit.edu/papers/pdf/2007_Calabrese_Colonna_Lovisolo_Parata_Ratti_senselab.pdf.

[18] Subbarao V. Wunnava, *et.al.*, "Travel time estimation using cell phones for highways and roadways," www.dot.state.fl.us/research-center/Completed_Proj/Summary_TE/FDOT_BD015_12_rpt.pdf

[19] Murat Ali Bayir, Murat Demirbas, Nathan Eagle, "Mobility Profiler: A Framework for Discovering Mobile User Profiles", technical report, 2008.

[20] www.wirelessintelligence.com.

[21] M Garofalakis, R Rastogi, K Shim, "SPIRIT: Sequential pattern mining with regular expression constraints," in Proceedings of VLDB 99, 1999.

[22] J Ayres, J Flannick, J Gehrke, T Yiu, "Sequential pattern mining using a bitmap representation," international Conference on Knowledge Discovery and Data Mining (KDD99'), pp. 429 - 435, 2002.

[23] J. Koolwaaij, A. Tarlano, M. Luther, A. Battestini, P. Nurmi, R. Vaidya, and B. Mrohs, "Context Watcher", http://www.lab.telin.nl/ oolwaaij/ showcase/crf/cw.html, 2005.

[24] J. Koolwaaij, A. Tarlano, M. Luther, P. Nurmi, B. Mrohs, A. Battestini, and R. Vaidya, "Context Watcher: Sharing context information in everyday life," in Proceedings of the IASTED conference on Web Technologies, Applications and Services (WTAS). IASTED, 2006.

[25] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams, "Mobile Landscapes: using location data from cell phones for urban analysis," Environment and Planning B: Planning and Design, vol. 33, pp. 727 - 748, 2006.

[26] Shannon, Information theory, 1948.

[27] Lan Wang, Zhisheng Niu, Yanfeng Zhu et.al., "Integration of SNR, Load and Time in Handoff Initiation for Wireless LAN," in proceeding of PIMRC'03, vol. 3, pp. 2032-2036, 7-10 Sept. 2003.

[28] Mobile Quality Analyzer, https://intouch1.p3-solutions.de/panelmanager/help/welcome_to_mqa.html.

[29] www.celltrack.com.

[30] http://www.cs.helsinki.fi/group/context/.

[31] Theodore S. Rappaport, Wireless communications principles and practice (2nd Edition), ISBN: 0130422320, Jan. 2002.

[32] George K. Zipf, *Human Behavior and the Principle of Least-Effort,* Addison-Wesley, 1949.