Precise Location Acquisition of Mobility Data Using Cell-id

Shafqat Ali Shad[1], Enhong Chen[2]

**[1] Laboratory of Semantic Computing and Data Mining,
Department of Computer Science and Technology
University of Science and Technology of China
Huangshan Road, Hefei, 230027 Anhui, China**

**[2] Laboratory of Semantic Computing and Data Mining,
Department of Computer Science and Technology
University of Science and Technology of China
Huangshan Road, Hefei, 230027 Anhui, China**

Abstract

Cellular network data has become a hot source of study for extraction of user-mobility and spatio-temporal trends. Location binding in mobility data can be done through different methods like GPS, service provider assisted faux-GPS and Cell Global Identity (CGI). Among these Cell Global Identity is most inexpensive method and readily available solution for mobility extraction; however exact spatial extraction is somehow a problem in it. This paper presents the spatial extraction technique of mobile phone user raw data which carries the information like location information, proximity location and activity of subjects. This work mainly focuses on the data pre-processing methodology and technique to interpret the low level mobility data into high level mobility information using the designed clustering methodology and publically available Cell-IDs databases. Work proposed the semi-supervised strategy to derive the missing locations thorough the usage of semantic tag information and removal of spatial outliers for precise mobility profile building by adopting the LAC-clustering, which is a variant of a hierarchical clustering.

Keywords: *Mobility data, Spatial extraction, Trajectory mining, GSM network*

1. Introduction

Successful extraction of user mobility profile from raw data can open the door of wide range of communication system and applications. Dwell time extraction and significant location finding are two most important prediction parameters in ubiquitous communication and computation. Most of the spatial data based application follow the mixed model based on discrete location and continuous time. So location extraction is of far more important.

Mobile phone data now days became an extensive source of spatial information for the potential applications like advertisement [1, 2], early warning systems[3], city wide sensing [4], pollution exposure, route marking and tracking [5], traffic management, social networking and community finding [6]. Low level data log of mobile data is interpreted to the mobility information so that it can be used for the extraction of mobility profiling.

For location identification two kinds of technologies are developed indoor and outdoor. While in case of indoor, these are mostly short distance based like Bluetooth, RFID or infrared e.g. Active badge[7] and Active bat [8]. And along with this Wifi [9] can also be a source of geo location provider for analysis and findings, While in case of outdoor most popular are GPS, Assisted faux GPS and GSM. In case of GPS the availability of open space is important and need high power consumption for mobile users along with installation of dedicated receivers. So, most convenient and standardized solution available is GSM (Global system for mobile communication) as it does not require any extra equipment installation being a standard in second generation cellular networks. Methods used in GSM network for location acquisition are Assisted-GPS (A-GPS), Time difference of arrival (TDOA) and Enhanced observed time difference (EOTD) but these all methods need upgrading of mobile handsets and extra equipment installation in the network [10]. And one main problem is that mobile operators do not make the location information collected through these mentioned methods publically available as being critical for their own backbone applications.

So the only choice is to use the Cell Global Identity (CGI) property of GSM network which is easy and readily available for location prediction. Commonly it is meant to be Cell ID positioning which uniquely identify the mobile user connected to the corresponding cell in the GSM network. However this CGI consists of four main head of information i.e. Mobile country code (MCC) assigned to every country, Mobile network code (MNC) assigned to every operator , Location area code (LAC) created by operator for identification of Cells, Cell ID which is given

by the Cell to each user connected. So MCC, MNC, LAC, Cell Id provide the unique set of identification for location extraction of any user at particular time span.

As CGI is an approximate location of user which is supposed to be converted to latitude and longitude coordinates so that specific location of mobile user can be traced out, this is a trivial problem. Firstly there is continuous re-labeling and introduction of new CGI because of fast growth in cellular networks. Secondly the network topology is made hidden from the public by the operator so large set of information is hidden for spatial information extraction. Thirdly information retrieved from publically available Cell ID databases is likely to be examined for its precise results. Fourth and last problem is that CGI is 4 header set of information so determination of specific heads in CGI is also important for location extraction as per GSM network architecture.

The above stated problems are the source of noise and outliers in the data so that they are likely to be removed so that the exact results can be extracted from any given dataset for spatial extraction and its usage for potential applications. So the focus of our work is on the extraction of location information thorough GSM data, estimation of missing locations using the semantically tagged location information, prediction and removal of spatial outliers in the data.

2. Related work

Over the period of time mobility attracted many researchers due to of is potential applications in the real world so lot of work has been done for the retrieval of spatial information and its processing. Like GSM data is used for city wide sensing [11-15] in these work the basic emphasize was to develop the model for collecting and processing the data from privately held sensors, traffic monitoring system[16] where personal sensors like camera and mobile phones can be used for information capturing, social network and human behavior analysis mining [17-19] in these projects most important task is related to finding socially important places from the trajectory data and then finding the social relationships among the users. Similarly CarTel system is designed by Hull et al. [20] which works on the basis of installed GPS sensors and cameras on taxi through opportunistic message forwarding. While some of the cell based location data projects [21-24] are related to the opportunistic message forwarding services depends on the location visited by the user during the mobility.

Zanoozi et al., [23] proposed the cell residence time optimization on through the human mobility analysis while Li et al., [25] proposed a prediction model for cell handover residence time on the basis of Markov model and Kalman filter to predict the next visit of user at a particular location. Akyildiz et al., [26] and Cayirci et al., [27]proposed a user visit prediction model using the direction of motion, speed, current position and history visits for network performance framework designing.

Most important work is done by Musolesi et al., [19] in which mobility models are divided into two categories i.e. called traces and synthetic models; the second modeling is more easily and commonly used because of difficult public data gathering. Gonzalez et al., [28] used 100k mobility data for the pattern analysis of mobile users based on cell location, his work presents the spatial-temporal nature of mobility where significance of locations is calculated and top K-locations are retrieved. Nurmi et al., [29]proposed a clustering technique for finding the significant places from the mobile user mobility traces. They used graph based transition modeling for cell towers network and cell location data for prediction. While Ficek et al., [30]used the Google API for the location abstraction but their work is more focused on the mobility finding rather missing value retrieval.

All of the above mentioned work is related to the mobility analysis but there is no significant work is carried out related to the data pre-processing on the raw dataset. Even if the pre-processing is carried out it is done on the human mobility data carrying the information like duration of stay, time of visit, density etc and most importantly the semantic tagging by the user is initially ignored for pre-processing of the information, while in our work we proposed the methodology to pre-process the raw location data after the deep study of GSM network so that the collected data become cleaned, ready for mobility analysis and outlier free. We considered the semantic tagging by the user over a particular location as a significant parameter beside the salient features of any GSM network. We used the clustering methodologies for data pre-processing by clustering the tagged location and adopting the LAC based methodology [30] for clustering the Location area in a GSM network, this work is focused only on pre-processing for retrieval of missing location values, outlier detection and availability of data for future mobility profile building. Though the proposed methodology is effective for cell oscillation resolution and significant place identification but we will not discuss here as mobility profiling is next task in our framework.

3. Methodology

3.1. CGI information headers

As mentioned the Reality data mining dataset has partial information of CGI i.e. LAC and Cell ID, so it is of obvious importance to determine whether this information is enough for retrieval of approximate location for potential usage. To understand this we must analyze the GSM network architecture. Due to of space limitation we are not providing detailed analysis but main features would be enlisted here.

Every mobile operator uses its own topology which is hidden from the public usage. To understand the topology of the GSM network operator let us have an introduction to the basics of it. The basic units of any cellular network is Base Transceiver Station (BTS), each BTS creates its own cell which is sectorized to allocate multiple Cell IDs within one cell.

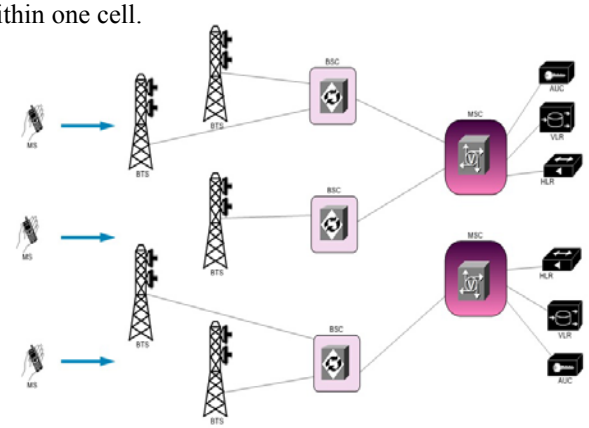

Fig. 1 Architecture of functional GSM network.

As shown in Fig 1. Mobile Station (MS) is connected with the Base Station Controller (BSC) and Mobile Services Switching Center (MSC). The MSC made link between different MSCs and BSCs to provide communication channel between two MS registered to two different cells.

Number of BTS determines the number of radio cells which are equal. This is of due importance that BTS connected to a MSC share same Location Area (LA). One or more MSCs together define a region identified by unique number called Location Area Code (LAC) as shown in Fig 2.

Fig. 2 Network topology: hexagonal view.

Each operator has different setup topology for BSCs connected with MSC, result in different number of Cell IDs in same LAC for each operator. Infact user density determines the LAC size, in urban areas LAC has few Cell IDs while in case of rural areas its much higher. If the number of users connected with MSC goes beyond its capacity this will make a hand shake with new MSC and handover upcoming users. So any GSM cell is uniquely identified by Cell ID in combination with LAC and Country Code. However LAC code is most important for the identification of geo-location in GSM network.

One main characteristic of GSM network is that the cells in GSM network are not of polygon shape, infact they are overlapped because of coverage and tradeoff issues in network. This overlapping is denser in urban areas while in rural areas it can long to Km. But can not be more then the 35Km which is a limitation parameter in any GSM network. This means that even a user is stationary it can be allocated different Cell IDs over time period. This is done due to of overlapping and this switching is very fast which appears to be a jump from one cell to another however it can not be taken as the movement of the user. This phenomenon is known as cell oscillation.

It is of due importance that LAC in any GSM network is organized in way that it can provide all the services available through GSM network to every connected mobile station. This depends on number of user density in area, messaging service, GPRS service; value added services and many other parameters. This LAC is quite dense in the urban areas where cells are likely to be overlapped for load balancing and quality of services, while in rural area this can be wider but not more than the 35km where user density is less. So it is equally useful to utilize the CGI information i.e. LAC and Cell ID for the retrieval of geo-location as a latitude and longitude corresponding to every Cell ID through publically available Cell Id databases.

3.2. Use of semantic information for missing values

Reality mining dataset carries the semantic information of the location as well as a human readable tags e.g. Home, Lab, Club, Airport etc. The main assumption in this regard is that even a location is semantically tagged by the user as significant place it can have different Cell IDs every time user visit it, as discussed before in section 3.1 because of overlapping cells oscillation phenomena. This determines that nearby cells can be clustered together and can be assigned to a single Semantic location. This cluster information can be used for exact spatial information extraction beside resolution of cell tower oscillation

problem (This is beyond the topic of this paper however this is a part of our Mobility profile building project).
We have developed a semantic based clustering algorithm to resolve this problem. Let us define the set of all locations to be L and set of all cells to be C. We define two mappings Ø and Ψ. Ø:L→C such that Ø(lk)={ ci ϵ C | ci contains lk } where lk ϵL whereas Ψ:C→L such that Ψ (cj)={ lp ϵ L | lp is in cj }. Now we choose a location say l_q and we calculate Ø(l_q)={c_i^q}. Next for every c_i^q, we calculate Ψ (c_i^q)={l_i^q}. we sort c_i^q's according to |Ψ (c_j)| and then store this sorted data aginst l_q. we also store |Ø(l_q)| against l_q. The pseudo code of the algorithms is as:

Algorithm: Semantic based location clustering
Select the complete list of the semantically tagged locations from the subject data
Treat each semantic location as a single cluster and repeat step 1-5 for each semantic location

1. Select all the Cell IDs associated with the selected semantic location
2. Compute the frequency of each Cell ID as per its appearance in the semantically tagged data
3. Sort the Cell IDs in descending order on the basis of their frequency
4. Compute the number of Cell IDs associated with the semantic location
5. Mark the semantic location as a single cluster which uniquely identifies all associated Cell IDs.

The main reason to cluster the semantically tagged location is to make the data consistent for the mobility profile building as this can resolve two main problems i.e.

3.2.1. Location extraction for missing cells

Any mobility data collected grows older over time due to of fast growth in GSM network, many cells are redefined as a result they can not be retrieved from the publically available Cell ID databases, so these missing Cell IDs can be cross checked from the clusters generated through the above algorithm. If these missing Cells are present in the cluster they can be replaced with their neighboring Cell IDs for the extraction of location along with their semantic tagging easily.

3.2.2. Cell switching and oscillation resolution

As discussed earlier in section 3.1 the cell oscillation is trivial problem for determination of significant place during the mobility profile building, even a stationary mobile station can be miss treated as moving because of Cell ID switching. However this problem can be resolved through above mentioned clustering technique, for example if there is rapid change in the Cell IDs over a certain period of time in the mobility path of the subject we can determine the sequence of oscillation pairs from it and then it can easily be checked whether in any of the generated cluster these pairs fall under same semantically tagged location or not, if they fall under same location this means the subject is stationary (Due to of topic restriction cell oscillation is not discussed in detail here).

3.2.3. Significance of location

During the mobility path building process the determination of significance parameter is an important task, however in the mentioned algorithm it is done somehow by assigning the frequency of any Cell ID in the data. For example if the subject is a student and spend most of his time in the school where most of building are in vicinity of each other like Laboratory, Canteen, Play ground there is great chance that one Cell ID can appear for multiple significant locations due to of overlapped cells. The frequency computed in the step 2 of the algorithm can be used for determination of significant places during mobility profile building. While the step 4 is also important as it will allow us to determine the number of cells associated with this location, for example if a Cell ID is present in more than one cluster we can choose a cluster with high number of location identifier during profile building process because it is more dense location.

3.3. Removal of spatial outliers

As mentioned before due of fast growth in GSM network many of the cells are reused and due to of its spatial nature, data collected can have spatial outlier which can lead to irregular trajectory or mobility profile building. For the removal of such outliers we can use the spatial clustering technique [30]. As the data is Spatial in nature and dense, it is different from the non-spatial data so we can not apply typical clustering techniques on it. However spatial data has one most important parameter i.e. distance attribute which can be used for its relative position determination in the space between neighboring points [31]. Over the period of time many clustering algorithms are developed for removal of outliers [32] but most of these algorithms are shape dependent or need user defined parameters.

As discussed in the section 3.1 location area is most important part of any GSM network as this is planned by the operator in such a way that density of Mobile station is covered equally and efficiently. This planning is dependent on the density of mobile stations, text messages, GPRS and other services [33].

While in case of the dense data it is recommended to use Euclidean distance for similarity measure which quite acceptable in the given situation where data is dense and spatial in nature with an obvious attribute of distance[34]. As most of the previous work done in this regard is dependent on the user behavior or trajectory data which provide the additional determination parameters like time of stay, position change, signal power measurements etc [35-38], our work is different in this regard that we are removing spatial outliers and detecting them by adopting the LAC clustering proposed by their work [30]. Let C= $\{c_1, c_2, \ldots, c_n\}$ be the set of cell id's in a specific location area. First we calculate Euclidean distance of all possible pairs of cell ids. Then we define a proximity matrix $P=[d_{ij}]$ where $1 \leq i, j \leq n$ and d_{ij} is the distance between c_i and c_j. for every $k \in \{1,2,\ldots,n\}$, spot out the smallest value of d_{kj}. Merge c_k and c_j. Now again calculate Euclidean distance and proximity matrix and repeat the steps mentioned above. Continue this process until we get a single cluster with maximum merging and assign it to the location area as representative. The pseudo code of algorithm is as:

Algorithm: Locality based clustering algorithm
1.1 Select complete list of unique location area codes from the subject data
1.2 Consider every location area code as a cluster and repeat the step 1-7 for every location area

1. List all the cell ids under the specified location area
2. Calculate the Euclidean distance between cell ids for similarity measure
3. Consider each cell id as a sub-cluster
4. Repeat until single sub-cluster with maximum locations is generated
5. Merge two similar sub- clusters depend on their distance measure
6. Update the Euclidean distance measurements for all sub-clusters under the location area
7. Use the distance parameter to define the clusters and pruning
8. Assign the sub cluster with highest location points to the location area as its representative

This algorithm described in a detailed manner [30] makes sure that all the locations represented by the location area are true, dense and outlier free. As each sub-cluster generated in step 4-6 is true representative of dense location within the location area, so at step 7-8 only single true most dense representative is assigned to the location area on every iteration. While the Euclidean distance determines that the location points linked or merged together are similar which ensures the removal of any spatial outlier form the data. This algorithm ensures that during the mobility profile building these clusters can be used as clean & outlier free set of unique locations (We are considering the profile building as our next step in our framework so it is out of scope of this work).

3.4. Google API and My Location (Beta) service

Open Cell ID database [39]with more than worldwide 620,000 unique cells, CellDB [40] with 180,000 cells and CellSpotting.com [41] with 111,000 cells are publically available databases for Cell IDs, however they differ in term of number of stored cells, data access methodology and content source. While Ericson Cell ID look-up API [42] and Location-API.com [43] are "crowd-sourced" based Cell ID databases, consist of around 7 Million Cell IDs each which are gathered thorough GPS. Main problem with these Cell ID databases is that they are sparse or have limited coverage. In this scenario the Google Cell ID database is more fit to any problem of Cell ID relative longitude and latitude coordinates retrieval.

Google inc. in 2007 went "crowd sourced" by introducing the Beta version of My location [44] service which works well enough for Cell Global Identity (CGI) information. This CGI information is sent to the Google API i.e. www.google.com/glm/mmap , which provides the longitude and latitude values of location to which CGI belongs to. Beside this location request method many work has been done regarding the direct use of Google API for the translation of CGI [45, 46] into latitude and longitude coordinates.

4. Dataset

Dataset for the proposed analysis is selected from Reality Mining Project group of MIT Media labs [47]. This data is collected over duration of 9 months involving the 100 people. Data is collected using the mobile set of Nokia 6600 with installed software to capture the continuous log recording of cell tower transitions along with other useful data.

As there is no GPS in the Nokia 6600 so only representation for location is Cell ID. However software provides the facility to the user to give a semantic name to a particular location through prompting. The total activity span is 350K hours and the size of database is approximately 1GB. This makes the Reality mining dataset a potential research source for extraction of mobility trends. But this data has partial location information i.e. MCC and MNC is not available for privacy reasons so it makes the dataset trivial to retrieve exact location of the user. However semantic tagging is available for some of the location by the users.

5. Experiments and results

As mentioned we have chosen the reality mining dataset for our experimental purposes. The results are as under.

5.1 Observing the semantic tagging by user

As discussed in previous sections our methodology depends on semantic tagging by the user to generate the missing values. We took the sample data of user X for this purpose. In reality mining dataset the semantic location information is present in ***s(n).cellnames*** table. We carried out the Algorithm 1 devised in section 3.2 on the table data and determined the Cells associated with the Semantic location and also computed their weight by calculating the number of cells associated. While on other hand we have calculated the number of location a particular Cell is holding to determine its significance in the dataset.

For user X we determined that total number of distinct locations is 75 which are used for the determination of cells associated with each location. During analysis we found that some locations are quite dense and are contains multiple cell ids associated with them so this information support our algorithmic assumption. For example in analysis we observed that locations "Home" and "Airport" are identified by 4 Cells each so these 4 Cells are in vicinity and can be associated to a single location cluster.

Later on if the coordinate value of one of them is not retrieved through the Cell Id i.e. Google Cell ID database we can use the coordinate value of the other Cells in same cluster for concise mobility representation. However the same set of cells can also be used for cell oscillation problem resolution, if there is rapid switching of cells at a particular span of time during mobility building of the user, we can check the switching pair whether it is the same pair of cells as appeared in the clusters or not, if it is same, it is obvious user is stationary and residing on the same location or semantically tagged location.

Table 1 shows information about some of the semantic locations tagged by the user along with the number of cells that are observed on it.

Table 1: Semantically tagged locations and cells incident

Semantic location	*Number of cells incidence*
Mgh	5
Office	4
Airport	4
Home	4
Greg's apt	4
Grand parents	3
Google	3
Redhat	3
Chicago O hare	2
Topohub	2

While Table 2 represents the some of the information about the particular cell that how many time it is been observed at different locations where each Cell represents LAC and Cell ID.

Table 2: Unique cell's appearance frequency

Cells (LAC.Cell ID)	*# of locations represented*
C30	7
C40	6
C23	6
C14	5
C56	4
C44	4
C25	3
C47	2
C27	2
C39	2

5.2 Location retrieval with Google API

As discussed earlier the reality mining dataset contains incomplete information i.e. MCC and MNC is missing in the locations headers, So it is only LAC and Cell ID which can be used for the location tracking. However during our experiment we observed that LAC and Cell ID is enough to collect the location information as Google Cell ID database is crowd source based and LAC is uniquely distributed among the operators in one country (As

discussed in section 3.1). Initially the total number of unique cell Ids against the user X can be determined from the table s(n).all_locs, which are 1744 for the subject under observation. When the Google cell Id database was requested thorough the batch tool developed in python using the Google Location APIs to fetch the location information, 680 cells are retrieved. Which are about 39% of the total number of unique cells observed against the user X. As the GSM network is changed over the time and most of the LAC are relabeled or reused resulting in change of Cell Ids and non-retrieval of location information. One main reason of this missing information is development of 3G networks also across the USA over last years and the reality mining dataset is almost 7 years old. However our assumption made in section 3.1 about the GSM network architecture is dealt well enough because we retrieved the location information only using the LAC and Cell ID as LAC is most important unit in any GSM network. After retrieving this information we implemented our semantic tagged based algorithm presented in section 3.2 on the missing values (in our case Google returned (0,0) information against these set of values), we retrieved the location information of additional 88 cells so total number of cells retrieved with location information are 768 which are 42% of the uniquely observed cells against user X. These results support our assumption made in section 3.2 that the semantic location information can be used for missing values retrieval. This retrieval ratio directly corresponds to the number of tagged locations if the number of tagged location is high there is high possibility of retrieval of missing locations, which is quite low in our observed case i.e. 75. The table 3 represents the consolidated information regarding the observations made during experiment.

Table 3: Unique cell's appearance frequency

	# of cells	*%*
Total number of unique cells against subject X	1744	100%
# of Cell's location retrieved thorough Google API	680	39%
# of Cell's location retrieved thorough Semantic	88	3%
Total# of Cell's location retrieved	768	42%

5.3 Removal of spatial outliers:

As discussed and presented in section 3.3 we adopted the proposed algorithm on the retrieved cells to remove the outliers from the dataset. As per our assumption we declared the minimum threshold of 10 Cells in one location to consider it as a cluster. As the basic similarity measure is Euclidian distance it worked well enough for dense mobility data regardless of its limitation over high distant objects for plane surface as the distance between cells in one LAC is not too high. So after applying the LAC clustering algorithm finally retrieved the 698 cells which are 40% of total unique cells visited by the user.

Figure 3 and Figure 4 shows the spatial extraction of the retrieved cells on Google map. The map shows the distribution of location over many regions on the map not limited to USA only. However these locations are thickly populated or dense in some areas in USA like.

But these 698 cells are cleaned and there are no spatial outliers in it. We observed two false negative with same LAC because of missing MNC and MCC information in it along with expected change in GSM network for up gradation and re arrangement.

Fig. 3 Unique cell's Location information retrieved from Google MAPs.

.

Fig. 4 Unique cell's Location information retrieved after clustering from Google MAPs.

The raw location information extracted from the Google API is plotted in Fig. 5 this shows the spatial outliers quite clearly in it where the data is distributed in some places unevenly. However in Fig. 6 the data is plotted after clustering where spatial outliers are removed from the data using the clustering algorithms. The figure shows the consolidated and more denser form of the same data where points are not scattered from the center. The figure 5 and 6 shows that the spatial outliers are obvious in the GSM data and can lead to inconsistent results while mobility profile building. Figure 6 shows only those cells which are clustered well enough to a specific location area. However each location area is thickly populated and carries immense cells to be considered as valid location. While Fig.7 shows the complete clustering results where green spots represents the clustered locations and red spots represents the outliers which are removed through the clustering algorithm. These clustered locations ensure the concise and complete mobility profile building in later stage.

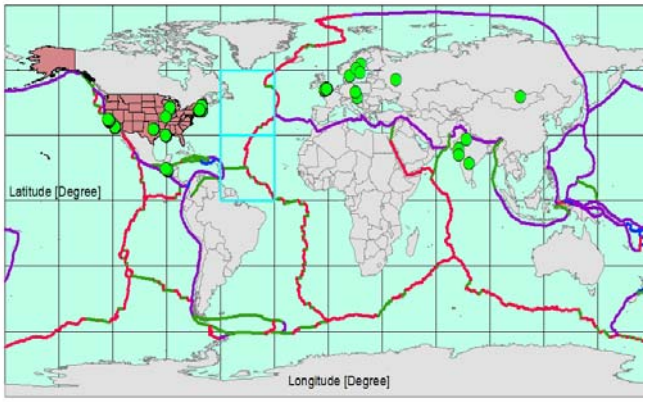

Fig. 5 Unique cells plot before clustering with expected spatial outliers.

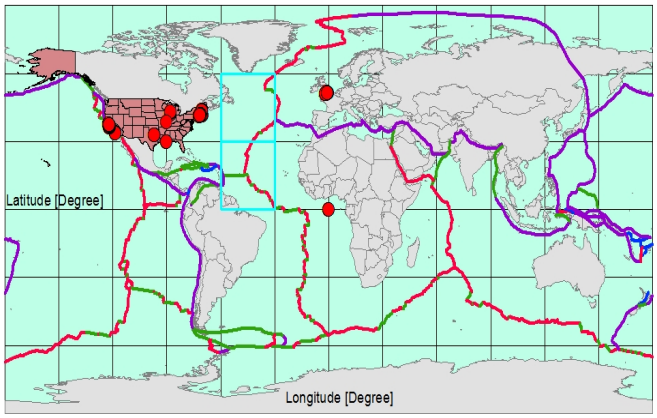

Fig. 6 Unique cells plot after clustering without of spatial outliers.

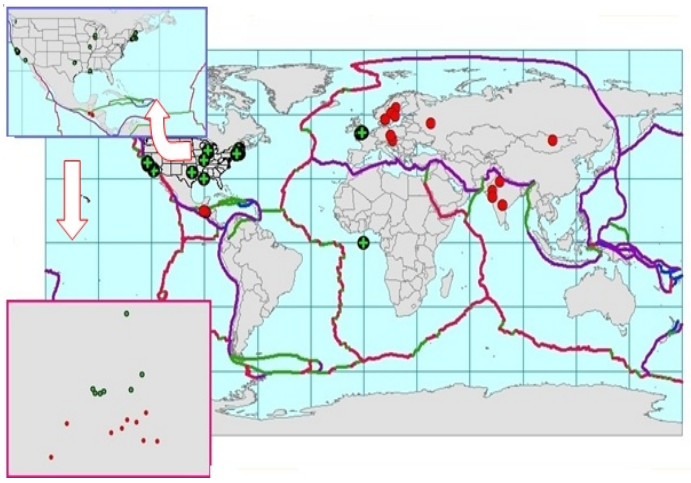

Fig. 7 Clustered cells result.

Figure 8. and Figure 9. Shows the complete comparison of the Cells retrieved through the experiments as Fig 8. Shows all unique cells retrieved through the Google API with expected outliers where each color represents the unique location area and its quite clear that there are certain points which are unequally scattered over coordinates regardless of their location area class, while in case of Fig 9. Where the Cells are retrieved after clustering they are dense, tightly bound to their location area class and outlier free.

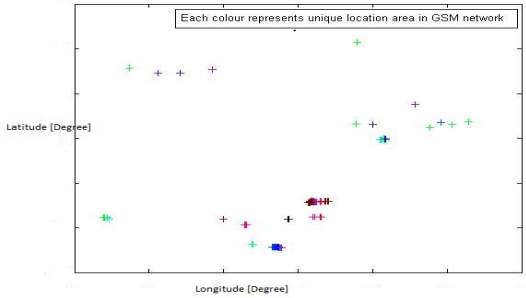

Fig. 8 All unique cells.

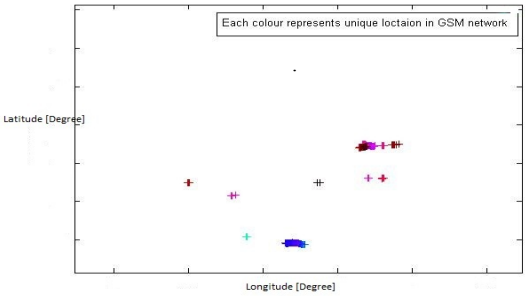

Fig. 9 Clustered cells.

We have computed the Hausdorff Distance which well known used in shape similarity measurements, we infact used Directional Hausdorff Distance in our experiment to find out the exact results regarding covered area and shape

similarity. Where Hausdorff Distance calculate the distance between two sets of points, X and Y (which could be two trajectories) in two dimensions in our case these are matrices X (longitude; latitude values for all unique cells) and Y (longitude; latitude values for unique cells clustered together). The Directional Hausdorff Distance (dhd) is defined as: dhd(X,Y) = max x c X[min y c Y [||x-y||]]. Intuitively dhd finds the point x from the set X that is farthest from any point in Y and measures the distance from x to its nearest neighbor in Y. So the Hausdorff Distance is defined as max{dhd(X,Y),dhd(Y,X)}, D is the matrix of distances where D(n,m) is the distance of the nth point in X from the mth point in Y. After retrieving the matrix containing the Hausdorff Distance measurements of two sets we plotted it on its axis. The Fig 11. Shows the plot result for the matrix which si again a straight line that shows that area covered by two point sets are similar and reliable so the set of coordinates retrieved through the clustering of Cell Ids are more reliable in term of mobility profile building. Additionally they are outlier free and compact in nature with more concise area coverage.

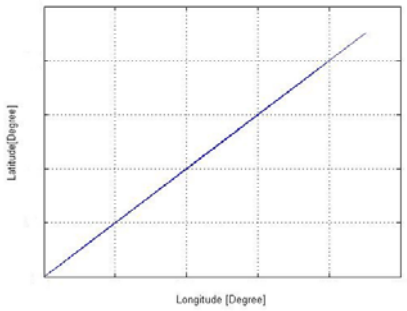

Fig. 11 Hausdorff distance matrix plot.

6. Conclusion and future work

In this paper our work mainly focused on the data pre-processing so that spatial information can be collected efficiently and precisely from the raw GSM mobility data without exploiting the user activity data which is used in most of previous works mentioned earlier. We resolved three main problem i.e. location information extractions by using semantic tag information, spatial outlier removal and partial resolution of cell oscillation problem using the raw GSM data. This resulting data offers possible investigation of interesting human behaviors i.e. mobility profile building, social network analysis, geographical proximity detection, route prediction.

In our future work we will work on Cell oscillation problem to extract the most significant places from the user activity data. After extracting these significant places we are interested to build up the user mobility profile over period of time for possible usage of user similarity measurements. As the data contains the location information, activity information and time stamps so this data can be used for all kind of potential mobility applications.

Acknowledgments

The work described in this paper was supported by grants from Natural Science Foundation of China (Grant No. 60775037), The National Major Special Science & Technology Projects (Grant No. 2011ZX04016-071), The HeGaoJi National Major Special Science & Technology Projects (Grant No. 2012ZX01029001-002) and Research Fund for the Doctoral Program of Higher Education of China (20093402110017,20113402110024).

References

1. Manuel, C., et al., Data Stream Processing on Real-time Mobile Advertisement :Ericsson Research Approach. 12th IEEE International Conference on Mobile Data Management, 2011. 1: p. 313 - 320
2.http://prodcat.ericsson.se/frontend/category.action?code=FGD%20101%20008, (downloded as on 3/4/2012).
3. Changhong, Y., et al., Use of Mobile Phones in an Emergency Reporting System for Infectious Disease Surveillance after the Sichuan Earthquake in China. Bulletin of the World Health Organization, 2009. 87(8): p. 619-623.
4. Kanhere, S.S., Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces Mobile Data Management (MDM) 12th IEEE International Conference, 2011. 2: p. 3-6.
5. Ling, C., et al., A personal route prediction system based on trajectory data mining. Information Sciences, 2011. 181(7): p. 1264-1284.
6. http://thenextweb.com/mobile/2011/07/03/the-rise-of-the-mobile-social-network/, downloaded as on 3/04/2012.
7. Harter, A. and A. Hopper, A distributed location system for the active office. IEEE Network, 1994. 8: p. 62-70.
8. Harter, A., et al., The Anatomy of a Context-Aware Application. Wireless Networks, 2002. 8: p. 187-197.
9. IEEE computer Society LAN MAN Standards Committee, W.L.M.A.C.M.a.P.L.P.S., IEEE Std 802.111 1999. The Institute of Electrical and Electronics Engineers, 1999.
10. "Location Technologies for GSM, G.a.W.N., White Paper, SnapTrack, A QUALCOMM Company, November 4, 2001.
11. Garetto, M. and E. Leonardi, Analysis of random mobility models with pde's. MobiHoc'06, 2006: p. 73–84.
12. T. Abdelzaher et al., Mobiscopes for Human Spaces. IEEE Pervasive Computing, 2007. 6(2): p. 20-29.
13. Demirbas, M., et al., iMAP: Indirect Measurement of Air Pollution with Cellphones. CSE Technical Reports, 2008.

14. Pentland, A., Automatic mapping and modeling of human networks. Physica A: Statistical Mechanics and its Applications, 2006. 378(41): p. 59-67.
15. Krause, A., et al., Toward community sensing. IPSN '08 Proceedings of the 7th international conference on Information processing in sensor networks 2008: p. 481–492.
16. K. Laasonen., Clustering and prediction of mobile user routes from cellular data. In PKDD, 2005: p. 569–576.
17. Lindgren, A., C. Diot, and J. Scott, Impact of communication infrastructure on forwarding in pocket switched networks. In SIGCOMM '06, 2006: p. 261–268.
18. L. Wang, Y. Jia, and W. Han, Instant message clustering based on extended vector space model. In ISICA, 2007: p. 435–443.
19. C.M.MircoMusolesi.Mobilitymodels for systems evaluation a survey. Survey 2008: p. downloaded as on 3/4/2012.
20. B. Hull et al., Cartel: adistributedmobile sensor computing system. In SenSys, 2006: p. 125–138.
21. Burke Jeff, et al., Participatory sensing. In ACM Sensys World Sensor Web Workshop, 2006.
22. Kirovski, D., et al., Health-os: A position paper. In ACM SIGMOBILE international workshop on Systems and networking support for healthcare and assisted living environments, 2007: p. 76-78.
23. Zonoozi, M. and P. Dassanayake, User mobility modeling and characterization of mobility pattern. IEEE Journal on Selected Areas in Communication, 1997. 15(7): p. 1239–1252.
24. E. Daly and M. Haahr, Social network analysis for routing in disconnected delay-tolerant manets. In MobiHoc, 2007: p. 32-40.
25. Markoulidakis, J., et al., Mobility modeling in third-generation mobile telecommunication systems. IEEE Personal Communications, 1997: p. 41-56.
26. Ian F. Akyildiz and W. Wang, The predictive user mobility profile framework for wireless multimedia networks. IEEE/ACM Trans. Netw, 2004. 12(6): p. 1021-1035
27. A. Chaintreau, et al., Impact of human mobility on opportunistic forwarding algorithms. IEEE Trans. Mob.Comput, 2007. 6(6): p. 606–620.
28. M. Gonzalez, C. Hidalgo, and A. Barabasi, Understanding individual human mobility patterns. Nature, 2008. 453(7196): p. 779–782.
29. Nurmi, P. and J. Koolwaaij, Identifying meaningful locations. In In Proc. 3rd Annual International Conference on Mobile and Ubiquitous Computing, IEEE Computer Society, 2006.
30. Ficek, M. and L. Kencl, "Spatial extension of the Reality Mining Dataset," Mobile Adhoc and Sensor Systems (MASS). IEEE 7th International Conference 2010: p. 666-673.
31. Shekhar, et al., Data mining for selective visualization of large spatial datasets. 14th IEEE international conference on tools with artificial intelligence (ICTAI), 2002.
32. Karahoc, A. and A. Kara, Comparing clustering techniques for telecom churn management. 5th WSEAS International Conference on Telecommunications and Informatics, 2006: p. 281-286.
33. http://www.tutorialspoint.com/gsm/gsm_addressing.htm: p. Downloaded as on 5/4/2012.
34. Zhongzhi.L and Xuegang.W, Spatial Clustering Algorithm Based on Hierarchical-Partition Tree. International Journal of Digital Content Technology and its Applications, 2010. 4(6): p. 210-215.
35. V. Bogorny, C. A. Heuser, and L. O. Alvares, A conceptual data model for trajectory data mining. In GIScience, 2010: p. 1-15.
36. P. O. V. de Melo, et al., Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. ECML PKDD: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010.
37. Miklos Kurucz, et al., Spectral clustering in telephone call graphs. 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis 2007: p. 82-91.
38. Farrahi, K. and D. Gatica-Perez, Discovering human routines from cell phone data with topic models. 12th IEEE International Symposium on Wearable Computers, 2008: p. 29-32.
39. 8motions. (2010) OpenCellId. [Online]. Available: http://www.opencellid.org/.
40. CellDB. (2008). [Online]. Available: http://celldb.org.
41. CellSpotting.com. [Online]. Available: http://www.cellspotting.com/.
42. Ericsson Labs. (2010) Mobile location. Available: https://labs.ericsson.com/apis/mobile-location.
43. Location-API.com. (2009) Combain Mobile AB. [Online]. Available: http://location-api.com/.
44. M. Chu. (2007) New magical blue circle on your map. Google Inc. [Online]. Available:http://googlemobile.blogspot.com/2007/11/newmagical-blue-circle-on-your-map.html.
45. Plusminus. (2007) Poor mans GPS –Cell (Tower)ID / Location Area Code - Lookup. [Online]. Available: http://www.anddev.org/map-tutorials-f18/poormans-gps-cell-tower-id-location-area-code-lookup-t257.html.
46. Acoustic. (2008) Learn How to Find GPS Location on Any SmartPhone, and then make it Relevant. Available: http://www.codeproject.com/KB/windows/DeepCast.aspx.
47. http://reality.media.mit.edu/download.php p. Downloaded as on 25/3/2012.

Shafqat Ali Shad: Mr. Shafqat Ali Shad is second year doctoral student at USTC. He obtained his Master degree in Computer Science from COMSATS Institute of Information Technology, Pakistan with the distinction of Chancellor Gold Medal in 2004. He had worked as National ICT Consultant for Asian Development Bank for Health sector reforms program in Pakistan mainly focused on Pakistan millennium development goals in 2010. He also worked with Planning Commission, Government of Pakistan as Deputy Director for ICT policy making and implementation of Five year plans, Annual plans and Vision 2030 from 2004 to 2009.

Prof. Enhong Chen: Dr. Chen Enhong, born in July 1968, currently works as a professor and doctoral supervisor at the Laboratory of Semantic Computing and Data Mining, University of Science and Technology of China (USTC). Prof. Chen is also a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE). Chen obtained his Ph.D. in Computer Software from USTC. Prof. Chen currently serves as Vice Dean of School of Computer Science and Technology of China,. Besides, Prof. Chen also serves on the program committees for over 20 international academic conferences. Prof. Chen has authored more than 90 research papers and invited papers published in international and domestic academic journals or submitted to international academic conferences. His paper presented to the top international conference on data mining, namely, the KDD2008, has won the Best Application Paper Award.