International Journal of Communication Networks and Security

Volume 1 | Issue 2 Article 6

July 2011

# An Efficient Approach Formulation of Social Groups of User Calls of GSM

Jaikumar M. Patil
*Department of CT, R.C.E.R.T., Chandrapur (M.S.), India*, jaipatil@yahoo.com

Rahila Patel
*R.C.E.R.T, Department of CT, R.C.E.R.T., Chandrapur (M.S.), India*, rahila.patel@gmail.com

# An Efficient Approach Formulation of Social Groups of User Calls of GSM

**Jaikumar M. Patil, Rahila Patel**
*R.C.E.R.T[1], Department of CT, R.C.E.R.T.,*
*Chandrapur (M.S.), India[1], Chandrapur (M.S.), India*
*E-mail : jaipatil@yahoo.com[1], rahila.patel@gmail.com*

***Abstract – We are living in a world of wireless technology. The most widely used wireless i.e. mobile computing device today is the Mobile phone, can be used not only for voice and data communications but also as a computing device running context aware applications. In this paper we present a model that based on GSM data base. The objective of this paper identifies social and suspicious groups based on Cell Id, IMSI, IMEI, date and time, Location Area, MCC and MNC. This information can be used by applications for the detection of users, user context, discovering of groups and relation between them using clustering technique of data mining. One of the vital means in dealing with these data is to classify or group them into a set of categories or clusters.***

***We demonstrate that even without knowledge of observed cell tower locations, we can recognize mobility modes that are useful for several application domains. Our mobility detection system was evaluated with GSM traces from the everyday lives of three data collectors.***

***Keywords – Social Groups, GSM, Clustering, Clustering algorithms, Data Sets, Model for Cluster Generation.***

## I. INTRODUCTION

In this paper our main objective an approach of finding user groups and their relations on the basis of GSM data base , extracting/generating real the data , transforming the data , processing the data and then formation of clusters using clustering algorithms for detecting fraud, detecting social groups etc. Social groups can be defined as sets of people, who have common interests, share their experience, express similar ways of thinking and reacting, share the same opinions, do similar things and have the same goals. There are various applications for social groups. For example, in marketing, if someone buys something, his family members and friends are likely to have the same interests to buy the same or a similar thing and have a similar level of income although we do not know how much they earn. So we may find potential buyers by social groups. Another important application for social groups is in the area of national security. For example, if somebody is a terrorist or robber, his intimate friends or socially close communication partners are likely (not necessary) to be terrorists or robbers too, since no law-abiding person wants to have some friends who are terrorists or robbers. One more application is used to quantify the telecommunication presence. On different days and at different times people usually would like to communicate with different groups of people. For example, we prefer to communicate with our colleagues in work time and to communicate with our family members, relatives and friends in non-work time.

GPS is one of the most popular positioning technologies because it provides world wide coverage, high accuracy, is free and a wide number of non-expensive GPS receivers are available. However, the main flaw of GPS technology is that it only works in open spaces where the satellites signal can be received, failing in narrow streets or inside buildings.

Our Personal Communication System allows mobile users to move from one location to another location since these systems are based on the notion of wireless access. In mobile system each mobile user is associated with Home Location Register (HLR) which stores up-to-date location of the mobile users. These logs accumulate as large database, in which data mining technique is applied to find the frequently followed locations [1].

## II. GLOBAL SYSTEM FOR MOBILE COMMUNICATION (GSM)

GSM is the standard for the second generation of mobile phones cellular networks, adopted by all the European countries. GSM is the most widespread cellular telephony standard in the world, with deployments in more than 400 countries. By the beginning of 2010, the GSM Association reported that there were over 4.0 billion users of GSM networks in the world [2]. A GSM base station is typically equipped with a number of directional antennas that define sectors of coverage, or cells. Each cell is allocated a number of physical channels based on the expected traffic load and the operator's requirements. Typically, the channels are allocated in a way that both increases coverage and reduces interference between cells.

Cell IDs are uniquely identified by the combination of Mobile Country Code (MCC), Mobile Network Code (MNC), Location Area Code (LAC), and cell id. Although other cell towers may be present in the area, our application only sees those associated with the phone's SIM card provider.

Cellular network is often represented as a set of joined hexagonal cells, which side-by-side covers one area, it is in fact made of cells with different and irregular shapes, with variable dimension, and that overlap partially. Identifying the cell that is being used by a mobile terminal is the most basic positioning mechanism for a GSM network. It is also the easiest to implement because this basic information is needed by the network core to keep track of the mobile phones and to be able to forward phone calls or deliver data. ITU assigns to each country a Mobile Country Code (MCC) and within each country a Mobile Network Code (MNC) is assigned to each cellular network operator [5]. Each operator is responsible for creating the Location Area Codes (LAC) for his network and to assign a numeric identification to each cell (cellID). Whenever a mobile terminal is connected to the network, it is associated

to one of these cells. Therefore, the absolute location of a terminal can then be expressed by the set {cellID, LAC, MNC, MCC}. Moreover, a mobile terminal can use networks other than its home network while in roaming. On the other hand, the cellID, LAC, MNC and MCC information of the serving cell is available at the mobile terminal, independently of the network or service provider

## III. CLUSTERING

The theme of clustering is to find individual groups, i.e. partition the data into groups so that most similar objects are to be found within the group and maximize the intra similarity values. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements.

There are basically several categories of clustering algorithms, namely, Partitioning algorithms, Hierarchical algorithms, Density based algorithms, Grid-based algorithms, Model based clustering algorithms, Ant based clustering algorithms and Genetic based algorithms. Partitioning algorithms like k-means divide data points into k number of clusters. K-means algorithm is a classical partitioned approach, provided with steady theory backgrounds. It is easy to understand and convenient for implementation. However, there are some disadvantages of k-means algorithm Hierarchical clustering naturally produces hierarchical trees instead of producing explicit clusters. Density based clustering algorithm defines clusters with high density regions and it discovers arbitrary shape clusters. Grid-based algorithms divide the space into regular cells in hierarchical structure and perform clustering operation on grid structure [5]. Genetic algorithms (GAs) are search and optimization procedures that are motivated by the principles of natural selection and natural genetics [7].

Well known algorithms in clustering are Zhan's Minimum Spanning Tree (MST) based clustering and clustering editing method, HCS algorithm, etc. Current research is focused on clustering using divide and conquers approach [5]. Usually this clustering methodology is used to detect irregular clustering boundaries in clustering results.

Clustering is used as a data processing technique in many different areas, including artificial intelligence, bioinformatics, biology, computer vision, city planning, data mining, data compression, earth quake studies, image analysis, image segmentation, information retrieval, machine learning, marketing, medicine, object recognition, pattern recognition, spatial database analysis, statistics and web mining.

The basic methodology or model used for the data representation is as follows:-

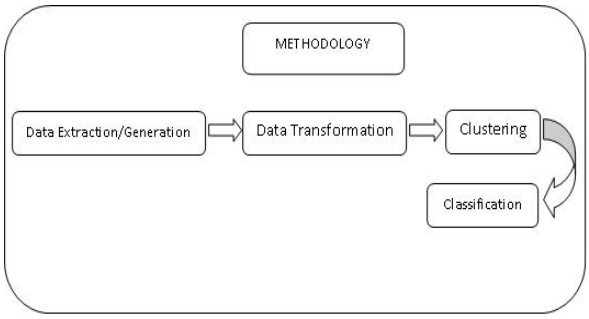


***Fig. 1 Data processing and classification model***

Clustering techniques are used for combining observed examples into clusters (groups) which satisfy two main criteria:
1. Each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
2. Each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

A large number of clustering algorithms have been developed to efficiently handle large size data sets. Most of these studies can be classified into four categories:

- Efficient Nearest Neighbour (NN) Search. One of the basic operations in any data clustering algorithm is to decide the cluster membership of each data point, which requires NN search. Algorithms for efficient NN search are either tree-based (e.g. kdtree [Moore, 1998, Muja & Lowe, 2009]) or random projection based (e.g., Locality Sensitive Hash [Buhler, 2001]).
- Data Summarization. Approaches in this category improve the clustering efficiency by first summarizing a large data set into a relatively small subset, and then applying the clustering algorithms to the summarized data set. Example algorithms include BIRCH [Zhang *et al.,* 1996], divide-and-conquer [Steinbach *et al.,* 2000], and coreset K-means [Har-peled & Mazumdar, 2004].
- Distributed Computing. Approaches in this category [Dhillon & Modha, 1999] divide each step of a data clustering algorithm into a number of procedures that can be computed independently. These independent computational procedures will then be carried out in parallel by different processors to reduce the overall computation time.

- Incremental Clustering. Approaches in this category [Bradley *et al.,* 1998] are designed to operate in a single pass over data points to improve the efficiency of data clustering. This is in contrast to most clustering algorithms that require multiple passes over data points before identifying the cluster centers.
- Sampling-based methods. Approaches in this category are algorithms like CURE [Guha *et al.,* 1998, Kollios *et al.,* 2003] that subsample a large dataset selectively, and perform clustering over the smaller set, which is later transferred to the larger dataset.

GENERAL TYPES OF CLUSTERS

*i) Well-separated clusters*
A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

ii) *Center-based clusters*
A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster.

*iii) Contiguous clusters*
A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

*iv) Density-based clusters*
A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

*v) Shared Property or Conceptual Clusters*
Find the clusters that share some common property or represent a particular concept.

*vi) Described by an Objective Function*
Find the clusters that minimize or maximize an objective function.

## IV. DATA SETS AND PARAMETERS

In this paper, the actual call logs are used for analysis. The actual call logs are collected or manually generated including their cellID, user IDs (Unique number representing a mobile phone user), date, time of calls, call direction (incoming and outgoing), incoming call description (missed, accepted), talk time, longitude, latitude and tower IDs (location of phone users). Based on the data collected or generated and then apply clustering algorithm for cluster formation. Then based on the clusters we divide social relationships into three categories: socially close members, socially near members and socially far members.

- ***Socially Close Members:*** These are the people with whom we maintain the strongest social relationship. Quantifying by phone calls we receive more calls from them and we tend to talk to them for longer periods of time. Family members, intimate friends and colleagues in the same team belong to this category.

- ***Socially Near Members:*** These relationships are not as strong as those of family members, intimate friends and colleagues in the same team. Sometimes, not always, we connect each other and talk for a considerably longer periods. We mostly observe intermittent frequency of calls from these people. Distant relatives, general friends, colleagues in a different team and neighbours are in this category.

- ***Socially Far Members:*** These people have weaker relationships with each other in social life. They call each other with less frequency. We seldom receive calls from them and talk each other in short time.

The affinity formula to classify social groups based on the time of the day, call frequencies, reciprocity and call duration. They have further defined the affinity A (P, Q) between P, Q:

- Socially close members if $0.7 < A(p,Q) \leq 1$

- Socially near members if $0.3 < A(p,Q) \leq 0.7$

- Socially far members if $0 < A(p,Q) \leq 0.3$

## V. MODEL FOR CLUSTER GENERATION

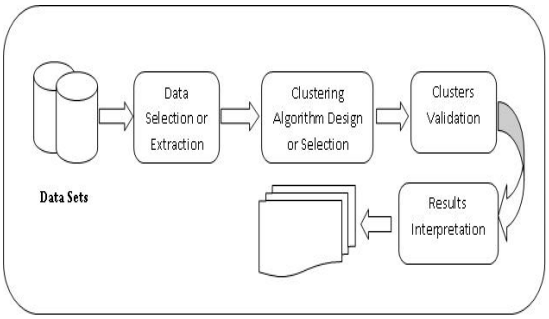


***Fig. 2 Creating Clusters using Data sets***

The Model that we used for the clusters formation as shown in Fig.2.Groups correspond to clusters of data. Cluster

analysis concerns a set of multivariate methods for grouping data variables into clusters of similar elements.

In this paper our basic and efficient approach based on the data or records of data set stored in the GSM data base or .csv file. The required data stored in the .csv file. First we extract the required data used for the formulate the user groups from the file then used some filtration technique for the standard format of data after that data transformation used for conversion of data into another form also remove errors and dissimilarities. Finally apply the clustering algorithms like K-means or K- Medoid is used for the cluster generation for the organizing the similar object or data and then grouping the data forms the groups. Classify the groups.

## VI.CONCLUSIONS

In this paper our main concentration on the approach of formulation of user groups and their relations on the basis of GSM data base , extract the data , processing the data , form the clusters and find structures using clustering algorithms in detecting fraud, detecting social groups etc. But if at all a group is to be traced which is not communicating with each other e.g., terrorists, then this systems will not work. An alternative model needs to be adopted which will cater to all the needs of finding user groups.

## REFERENCES


***[1] Sakthi, U., Bhuvaneswaran R.S. (2009): Mobility Prediction of Mobile Users in Mobile Environment Using Knowledge Grid, IJCSNS International Journal of Computer Science and Network Security Vol.9 No.1, 303-309.***

***[2] Miklós Kurucz ,András Benczúr, Károly Csalogány, László Lukács (2009): Spectral Clustering in Telephone Call Graphs, ACM 978-1-59593-848-0.***

***[3] Filipe Meneses, Adriano Moreira, Dept. of Information Systems University of Minho, Portugal: Using GSM CellID Positioning for Place Discovering.***

***[4] Pradeep Rai and Shubha Singh: A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010.***

***[5] K. Rajendra Prasad And Dr. P.Govinda Rajulu: A Survey On Clustering Technique For Datasets Using Efficient Graph Structures, International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2707-2714.***

***[6] O.A. Mohamed Jafar and R. Sivakumar: Ant-based Clustering Algorithms: A Brief Survey, International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010 1793-8201.***

***[7] Rahila H. Sheikh, M M Ranghuwanshi and Anil Jaiswal: Genetic Algorithm Based Clustering: A Survey, First International Conference on Emerging Trends in Engineering and Technology IEEE 2008.***

***[8] Hollmen, J., V. Tresp, and O. Simula (1999). A Self-organizing map for clustering probabilistic models Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN' 99), Volume 2, pp. 946-951 IEEE.***

***[9] Hollmen, J. and V. Tres, A hidden markov model for metric and event based data, Proceedings of EUSIPCO-X European Signal Processing Conference, Volume II, pp. 737-740.***

***[10] Huiqi Zhang and Ram Dantu.: Discovery of Social Groups using Call Detail Records, Dept of Computer Science and Engineering, Denton, USA.***

***[11] Ogatha, H. (2001): Computer Supported Social Networking for Augmenting Cooperation. Computer Supported Cooperative Work, 10, Kluwerr Academic Publishers 189-209.***

***[12] Kari,Laasonen : Clustering and Prediction of Mobile User Routes from Cellular Data, Dept. of C.S. University of Helsinki.***

***[13] Ning, Song (2006): Discovering User Context with Mobile Devices: Location and Time, Submitted for partial fulfilment of degree of M.E. at Massachusetts Institute of Technology.***

***[14] Paul McCarthy (2005): Forensic Analysis of Mobile Phones a thesis submitted for the Bachelor of Computer and Information Science (Honours) Degree.***

***[15] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE Survey of Clustering Algorithms ,IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.***

***[16] Anil K. Jain Department of Computer Science & Engineering Michigan State University East Lansing, Michigan 48824 USA : Data Clustering: 50 Years Beyond K-Means1.***

***[17] Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana and Yo-Ping Huang: Survey of Fraud Detection Techniques.***

***[18] Yoni De Mulder and George Danezis: Identification via Location-Profiling in GSM Networks.***

***[19] A.K. JAIN, M.N. MURTY and P.J. FLYNN: Data Clustering: A Review.***

***[20] P Burge, J Shawe-Taylor, C Cooke, Y Moreau, B Preneel, C Stoermann: Fraud Detection and Management in Mobile Telecommunications Networks.***

***[21] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE: Survey of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.***

***[22] Filipe Meneses, Adriano Moreira, Dept. of Information Systems, University of Minho: Using GSM CellID Positioning for Place Discovering.***