Department of Electrical Engineering
Indian Institute of Technology, Gandhinagar

Recursive Clustering of Mobile Phone Call Database B.Tech Project Report

Kishore Rathinavel
Department of Electrical Engineering
Indian Institute of Technology, Gandhinagar

Pawan Lingras
Department of Computer Science and Engineering
Indian Institute of Technology, Gandhinagar

May 8, 2012

Abstract

Phone call records of a group of customers can be a great source of information for user modeling with a view to glean customer behavior and preferences. Customer profiling is one of the preliminary and primary aspects of a customer relationship management strategy. The calling pattern can be useful in suggesting appropriate modifications to the customers calling plan. Extending the analysis for all the users can be useful for the phone companies to better manage their resources as well as to maximize utilization and profits. This project builds on earlier studies of mobile phone data mining to explore a novel technique of reiterative network mining. Using the techniques of crisp clustering previously developed, we explore the idea of applying the techniques reiteratively. The idea is that after each clustering, the data is modified slightly to retain the information about the clusters and further clustered repeatedly. We explect to see new patterns which were previously hidden and thereby advance the research in this field.

Keywords: Clustering, mobile call mining, recursive network mining, k-means

1 Acknowledgements

I would like to thank Prof. Pawan Lingras for his immense support and guidance in this project. I learnt a lot of things from his course and by doing this project under his guidance and I am immensely grateful to him. I would also like thank Adit Gupta and Lalitha Anusha for their help in some of the sections of this project.

2 Introduction

Mobile phones usage has topped five billion mark, with more than a billion new mobile phone connections added in 2009 and first half of 2010 [1]. The total number of mobile phones as a percentage of world population now stands at 72%, with ratios above 100% in many European regions. Mobile devices are going to be a dominant competitor to the personal computer based communication. Mobile communication devices are gaining even faster ac- ceptance than the proliferation of web in 1990s. Characterization of users is an important issue in the design and maintenance of mobile services with unprecedented commercial implications. Mobile phone data offers a fertile ground for data mining research as well as business analysis. In this project, we use a dataset created by Eagle, et al. [6] The dataset includes detailed logs of more than a 100 users' mobile phone activity over a nine month period. The logs are further supported with the help of descriptive surveys of all the users. Eagle, et al. [6] used the dataset for studying the social relationship between the users. We will be using the README file [5] provided for the data in order to understand it and use it.

This emerging area of application is called mobile phone call mining, which involves application of data mining techniques to discover usage patterns from the mobile phone call data.

The mining of mobile phone call datasets is gaining increasing attention from the research community. This section provides a brief review of some of the recent literature. The user models can be used for a number of applications in telecommunication industry including fraud detection, viral and targeted marketing, churn prediction [8]. Hohwald, et al. [8] report different applications of user modeling from large real-world datasets of mobile phone and landline subscribers. Their dataset consisted of six month phone records of 50,000 mobile phone users and 50,000 landline users from the same geographical area for a period of six months. The study presented aggregate behavior patterns and concluded that there are numerous differences between mobile phone and landline users that have relevant practical implications.

An unsupervised identification of customer profiles makes it possible for an organization to identify previously unknown characterizations of groups of customers. Clustering is one of the most frequently used unsupervised data mining technique. It is used at various stages in data mining from preliminary exploration of a new dataset, identification of outliers, as well as sophisticated analysis for decision making. Clustering has been used in a wide variety of applications from engineering [20, 11], web mining [9, 12, 14, 16] , to retail data mining [15]. This project applies the clustering technique to a new and emerging field of mobile call mining [6].

Researchers have so far used various methods of clustering such as fuzzy, crisp, kmeans, spectral clustering and granular clustering etc to mention a few. In this project we use a recursive clustering process to discover new patterns. The objective is to modify the data point details slightly after each clustering iteration. The clustering is done repeatedly over this data set.

We explore different clustering techniques applied in a recursive form to determine if the results are different. We believe the results be different and new patterns that were previously hidden will be discovered.

We use the popular clustering technique, namely, K-Means algorithm [7], [17] for crisp clustering. Other popular forms of clustering such as fuzzy C-means algorithm [4, 2] for fuzzy clustering will be done as part of this project in the future. Here, we discuss several clustering techniques that we intend to experiment with in the future. Conventional crisp clustering techniques categorize objects precisely into one group. The crisp clustering scheme is useful for a concise description of the clusters, while fuzzy clustering scheme is used for memberships of individual objects to different clusters, which is descriptive. However, in real-world applications, an object may exhibit characteristics from different groups. Soft clustering techniques such as fuzzy [4], [2] and rough clustering [14, 18] make it possible for an object to belong to multiple clustering leading to fuzzy or rough and overlapping boundary regions. We may use a procedure proposed by Joshi et al. [10] for creating a rough clustering scheme from the fuzzy clustering for concise comparison of cluster cardinalities, while still maintaining the overlapping nature of the fuzzy clustering scheme.

Datasets in many applications can be viewed at different levels of granularity. An information granule allows us to control the level of details that will be used in an analysis of a problem. Depending on the level of granularity, data mining techniques can produce different results. Correlating results from different levels of granularity can improve the quality of analysis. For the mobile phone dataset, we can analyze phone calls which are smaller or finer information granules. Since the phone numbers correspond to a coarser granule encapsulating a number of finer phone call granules, it will be helpful if the phone calls also corresponded to a logical grouping of phone numbers. Granular computing in general and granular clustering in particular has been an active area of research [19, 21]. Lingras, et al. [13] used the mobile phone calls to study relationship between clustering at two different levels of granularity of a dataset. Their objective was to identify the calling pattern of a customer. Their study used K-means algorithm and Davies-Bouldin (DB) index [3] along with cluster densities to compare the distance based quality of clustering schemes. However, the distance based clustering scheme does not necessarily provide an indication of semantics. We want to use the clustering scheme to make recommendations about the calling plans. With the help of Davies-Bouldin index we will attempt to recommend appropriate clustering schemes at both the levels of granularity.

3 Review of K-means clustering

Let X = $\{x_1, ..., x_n\}$ be a finite set of objects. Assume that the objects are represented by m-dimensional vectors. A clustering scheme groups n objects into k clusters C = $\{c_1, ..., c_k\}$. The name K-means originates from the means of the k clusters that are created from n objects. Let us assume that the objects are represented by m-dimensional vectors. The objective is to assign these n objects to k clusters. Each of the clusters is also represented by an m-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing k objects as the centroids of the k clusters. The objects are assigned to one of the k clusters based on the minimum value of the distance d(v, x) between the object vector v = $(v_1, ..., v_j, ..., v_m)$ and the cluster vector x = $(x_1, ..., x_j, ..., x_m)$. The distance d(v, x) is given by: d(v,x)=$\sqrt{\frac{\sum_{j=1}^{m}(v_j - x_j)^2}{m}}$ After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as: x$_j = \frac{\sum_{v \in x} v_j}{Size\ of\ cluster\ x}, where\ 1 \leq j \leq m$. The process stops when the centroids of clusters stabilize, i.e. the centroid vectors from the previous iteration are identical to those generated in the current iteration.

4 Data Preparation

This section describes data preparation applied to the original dataset provided by Eagle [5] followed by the design of the experiment.

The objective of the present study is to use recursive clustering to converge to a set of user profiles. The data set comprises of 182,208 phone calls data collected from about 102 users over a period of nine months. We chose the following six variables to represent a phone call:

1. Weekend/Weekday (1/0)
2. Daytime/night-time (1/0) (8 am - 7:59 pm was designated as daytime.)
3. Duration of the phone call (normalized using different weighting schemes)
4. Direction (outgoing/incoming = 1/0. Here missed call was considered an incoming call.)
5. Missed call (yes/no = 1/0. It was assumed that missed call was a special message. That is, the person did not want to talk to the caller. In some cases, missed call is used to deliver an agreed upon message without having to waste time or money through the connection.)
6. Voice call(1/0)
7. SMS(1/0)

8. Packet data(1/0)

9. MMS(1/0)

10. Data Call(1/0)

11. Long Duration call(1/0) This is a field that was added to the existing fields after the first clustering analysis of the phone calls. The need for this field will be justified later. If the duration is above 1512s, it is considered a long duration call.

Except for the duration of the call, all the variables had binary values while clustering phone calls. In the original data set, the type of call was classified as SMS, voice, data call, MMS, or packet data. MMS calls amount to just 25 calls. MMS calls and data calls amount to just 25 and 16 calls respectively. The dataset for the phone numbers is represented by the following variables:

1. Average duration of phone calls

2. Average number of Weekend/Weekday(1/0)

3. Average number of Daytime/night-time(1/0)

4. Average number of outgoing/incoming(1/0). Here, missed calls were considered as incoming calls.

5. Average number of missed calls(1/0)

6. Average number of SMS(1/0)

7. Average number of Voice calls(1/0)

8. Average number of long duration calls(1/0) This field was added after the first clustering analysis of the phone calls. The need for this field will be justified later.

5 Analysis of Raw data

We perform clustering on the phone calls dataset to glean some useful information. We ignore packet calls, MMS calls and data calls in this clustering. The durations are normalized with their mean. The normalized durations has a distribution as shown in the Table 1:

Minimun	1^{st} Quantile	Median	Mean	3^{rd} Quantile	Maximum
0.00	0.00	0.105	1.00	0.6202	129.00

Table 1: Distribution of duration of the phone calls

Since the maximum is disproportionately large compared to the mean and the 3^{rd} Quantile, we can infer the presence of some outliers. Upon performing K-means clustering on the calls, we get the optimum number of clusters to be 13. The cluster centers for the duration are shown in the Table 2:

Cluster Number:	1	2	3	4	5	6	7	8	9	10	11	12	13
Duration center:	0.18	97.76	0.07	64.88	23.14	7.89	16.64	44.73	2.72	4.91	11.72	1.18	32.04

Table 2: The duration parameter of the cluster centers

The cluster plots also suggest some average distribution and some outliers. Hence, it is required to split the data set into 2 sets of phone calls depending on the duration of the call. Observing the duration center of the clusters suggests that the duration can be split at some duration belonging to the clusters centered at 11.72 or 16.64. We find out the least duration in the cluster centered at 16.64 and use that as the duration to differentiate the low duration calls and the high duration calls. The least duration in cluster centered at 16.64 was 14s. But this is a normalized value which corresponds to 1512 in the original dataset. This gives us 2 new averages for phone call durations. The low calls duration have an average of 75s and the high duration calls have an average of 2500s. We use these new average values to normalize their respective groups. Henceforth, the clustering for the high duration calls and the low duration calls will be performed seperately. In order to represent the classification of the high and low duration calls, we now prepare a new dataset which has the field long duration as mentioned earlier. The long duration field is 1 if the duration exceeds 1512s otherwise, it is 0. When clustering is performed on the 2 datasets seperately, we observe that the clusters quality is very much improved.

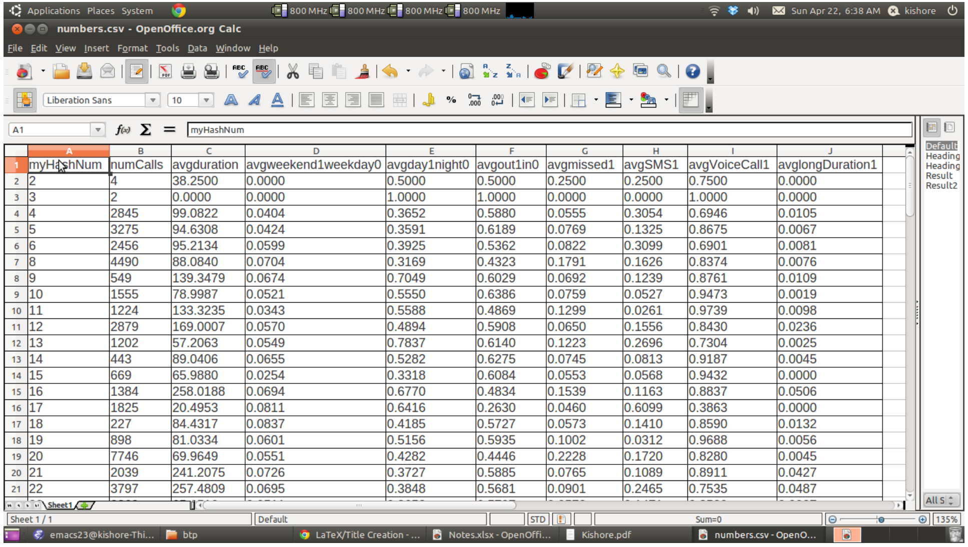
6 Design of experiment

In this paper, we attempt a recursive method of clustering where in each iteration we use some of the clustering information obtained in the previous iteration. The proposed method of carrying forward cluster information is the following:

1. Normalize the data by dividing the data by the mean of the corresponding values.
2. Perform the initial clustering analysis on the raw phone number data
3. Additional to the raw data, add columns for each cluster obtained in the previous step. With the data prepared in the previous section, we get the optimum number of clusters as 4 and hence we introduce 4 additional columns to the raw data set.

4. For each of the origin phone numbers, we have many destination phone numbers and we can group the destination numbers which are also origin numbers into groups based on the clustering of the origin phone numbers. To elucidate this clearly, consider that the clustering of phone numbers produces 4 clusters. For an origin phone number in cluster 1, we have a set of destination phone numbers. It is possible that some of the destination phone numbers also belong to cluster 1. Similarly, we will have destination phone numbers belonging to clusters 2, 3 and 4. We count the number of such destination phone numbers belonging to cluster 1, 2, 3 and 4 for each of the origin phone numbers. This new information is first normalized by their sum and entered in the columns we introduced in step 2. Screenshots of the dataset before and after addition of 4 columns is shown in Figures1 and 2.

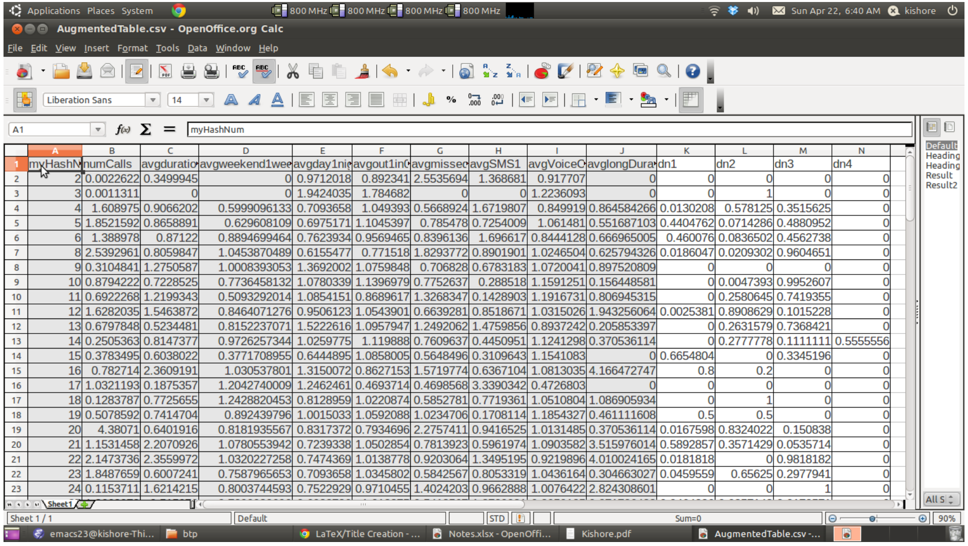
Figure 1: Before addition of the present cluster details


Applications Places System
800 MHz 800 MHz 800 MHz 800 MHz
Sun Apr 22, 6:38 AM kishore
numbers.csv - OpenOffice.org Calc
File Edit View Insert Format Tools Data Window Help
Liberation Sans
10
Default
Heading
Heading
Result
Result2
All S
Sheet1
Sheet 1 / 1
Default
STD
Sum=0
135%
emacs23@kishore-Thi...
btp
LaTeX/Title Creation - ...
Notes.xlsx - OpenOffi...
Kishore.pdf
numbers.csv - OpenO...

A1 = myHashNum

	A	B	C	D	E	F	G	H	I	J
1	myHashNum	numCalls	avgduration	avgweekend1weekday0	avgday1night0	avgout1in0	avgmissed1	avgSMS1	avgVoiceCall1	avglongDuration1
2	2	4	38.2500	0.0000	0.5000	0.5000	0.2500	0.2500	0.7500	0.0000
3	3	2	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000	1.0000	0.0000
4	4	2845	99.0822	0.0404	0.3652	0.5880	0.0555	0.3054	0.6946	0.0105
5	5	3275	94.6308	0.0424	0.3591	0.6189	0.0769	0.1325	0.8675	0.0067
6	6	2456	95.2134	0.0599	0.3925	0.5362	0.0822	0.3099	0.6901	0.0081
7	8	4490	88.0840	0.0704	0.3169	0.4323	0.1791	0.1626	0.8374	0.0076
8	9	549	139.3479	0.0674	0.7049	0.6029	0.0692	0.1239	0.8761	0.0109
9	10	1555	78.9987	0.0521	0.5550	0.6386	0.0759	0.0527	0.9473	0.0019
10	11	1224	133.3235	0.0343	0.5588	0.4869	0.1299	0.0261	0.9739	0.0098
11	12	2879	169.0007	0.0570	0.4894	0.5908	0.0650	0.1556	0.8430	0.0236
12	13	1202	57.2063	0.0549	0.7837	0.6140	0.1223	0.2696	0.7304	0.0025
13	14	443	89.0406	0.0655	0.5282	0.6275	0.0745	0.0813	0.9187	0.0045
14	15	669	65.9880	0.0254	0.3318	0.6084	0.0553	0.0568	0.9432	0.0000
15	16	1384	258.0188	0.0694	0.6770	0.4834	0.1539	0.1163	0.8837	0.0506
16	17	1825	20.4953	0.0811	0.6416	0.2630	0.0460	0.6099	0.3863	0.0000
17	18	227	84.4317	0.0837	0.4185	0.5727	0.0573	0.1410	0.8590	0.0132
18	19	898	81.0334	0.0601	0.5156	0.5935	0.1002	0.0312	0.9688	0.0056
19	20	7746	69.9649	0.0551	0.4282	0.4446	0.2228	0.1720	0.8280	0.0045
20	21	2039	241.2075	0.0726	0.3727	0.5885	0.0765	0.1089	0.8911	0.0427
21	22	3797	257.4809	0.0695	0.3848	0.5681	0.0901	0.2465	0.7535	0.0487

Figure 2: After addition of the present cluster details.



AugmentedTable.csv - OpenOffice.org Calc
File Edit View Insert Format Tools Data Window Help
Liberation Sans
14
A1
myHashNum
Sheet1
Sheet 1 / 1
Default
STD
Sum=0
90%

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	myHashN	numCalls	avgduratio	avgweekend1wee	avgday1nig	avgout1in0	avgmissed	avgSMS1	avgVoiceC	avglongDura	dn1	dn2	dn3	dn4
2	2	0.0022622	0.3499945	0	0.9712018	0.892341	2.5535694	1.368681	0.917707	0	0	0	0	0
3	3	0.0011311	0	0	1.9424035	1.784682	0	0	1.2236093	0	0	1	0	0
4	4	1.608975	0.9066202	0.5999096133	0.7093658	1.049393	0.5668924	1.6719807	0.849919	0.864584266	0.0130208	0.578125	0.3515625	0
5	5	1.8521592	0.8658891	0.629608109	0.6975171	1.1045397	0.785478	0.7254009	1.061481	0.551687103	0.4404762	0.0714286	0.4880952	0
6	6	1.388978	0.87122	0.8894699464	0.7623934	0.9569465	0.8396136	1.696617	0.8444128	0.666965005	0.460076	0.0836502	0.4562738	0
7	8	2.5392961	0.8059847	1.0453870489	0.6155477	0.771518	1.8293772	0.8901901	1.0246504	0.625794326	0.0186047	0.0209302	0.9604651	0
8	9	0.3104841	1.2750587	1.0008393053	1.3692002	1.0759848	0.706828	0.6783183	1.0720041	0.897520809	0	0	0	0
9	10	0.8794222	0.7228525	0.7736458132	1.0780339	1.1396979	0.7752637	0.288518	1.1591251	0.156448581	0	0.0047393	0.9952607	0
10	11	0.6922268	1.2199343	0.5093292014	1.0854151	0.8689617	1.3268347	0.1428903	1.1916731	0.806945315	0	0.2580645	0.7419355	0
11	12	1.6282035	1.5463872	0.8464071276	0.9506123	1.0543901	0.6639281	0.8518671	1.0315026	1.943256064	0.0025381	0.8908629	0.1015228	0
12	13	0.6797848	0.5234481	0.8152237071	1.5222616	1.0957947	1.2492062	1.4759856	0.8937242	0.205853397	0	0.2631579	0.7368421	0
13	14	0.2505363	0.8147377	0.9726257344	1.0259775	1.119888	0.7609637	0.4450951	1.1241298	0.370536114	0	0.2777778	0.1111111	0.5555556
14	15	0.3783495	0.6038022	0.3771708955	0.6444895	1.0858005	0.5648496	0.3109643	1.1541083	0	0.6654804	0	0.3345196	0
15	16	0.782714	2.3609191	1.030537801	1.3150072	0.8627153	1.5719774	0.6367104	1.0813035	4.166472747	0.8	0.2	0	0
16	17	1.0321193	0.1875357	1.2042740009	1.2462461	0.4693714	0.4698568	3.3390342	0.4726803	0	0	0	0	0
17	18	0.1283787	0.7725655	1.2428820453	0.8128959	1.0220874	0.5852781	0.7719361	1.0510804	1.086905934	0	1	0	0
18	19	0.5078592	0.7414704	0.892439796	1.0015033	1.0592088	1.0234706	0.1708114	1.1854327	0.461111608	0.5	0.5	0	0
19	20	4.38071	0.6401916	0.8181935567	0.8317372	0.7934696	2.2757411	0.9416525	1.0131485	0.370536114	0.0167598	0.8324022	0.150838	0
20	21	1.1531458	2.2070926	1.0780553942	0.7239338	1.0502854	0.7813923	0.5961974	1.0903582	3.515976014	0.5892857	0.3571429	0.0535714	0
21	22	2.1473736	2.3559972	1.0320227258	0.7474369	1.0138778	0.9203064	1.3495195	0.9219896	4.010024165	0.0181818	0	0.9818182	0
22	23	1.8487659	0.6007241	0.7587965653	0.7093658	1.0345802	0.5842567	0.8053319	1.0436164	0.304663027	0.0459559	0.65625	0.2977941	0
23	24	0.1153711	1.6214215	0.8003744593	0.7522929	0.9710455	1.4024203	0.9662888	1.0076422	2.824308601	0	0	1	0

5. We perform the next iteration of clustering by using the new data set we have prepared. We repeat the process from step 2 till the clusteirng results converge.

The above algorithm is explained diagramatically by the figures 3 and 4

Figure 3: This figure shows how static data is retained from iteration to iteration whereas the dynamic data is derived or calculated for each iteration

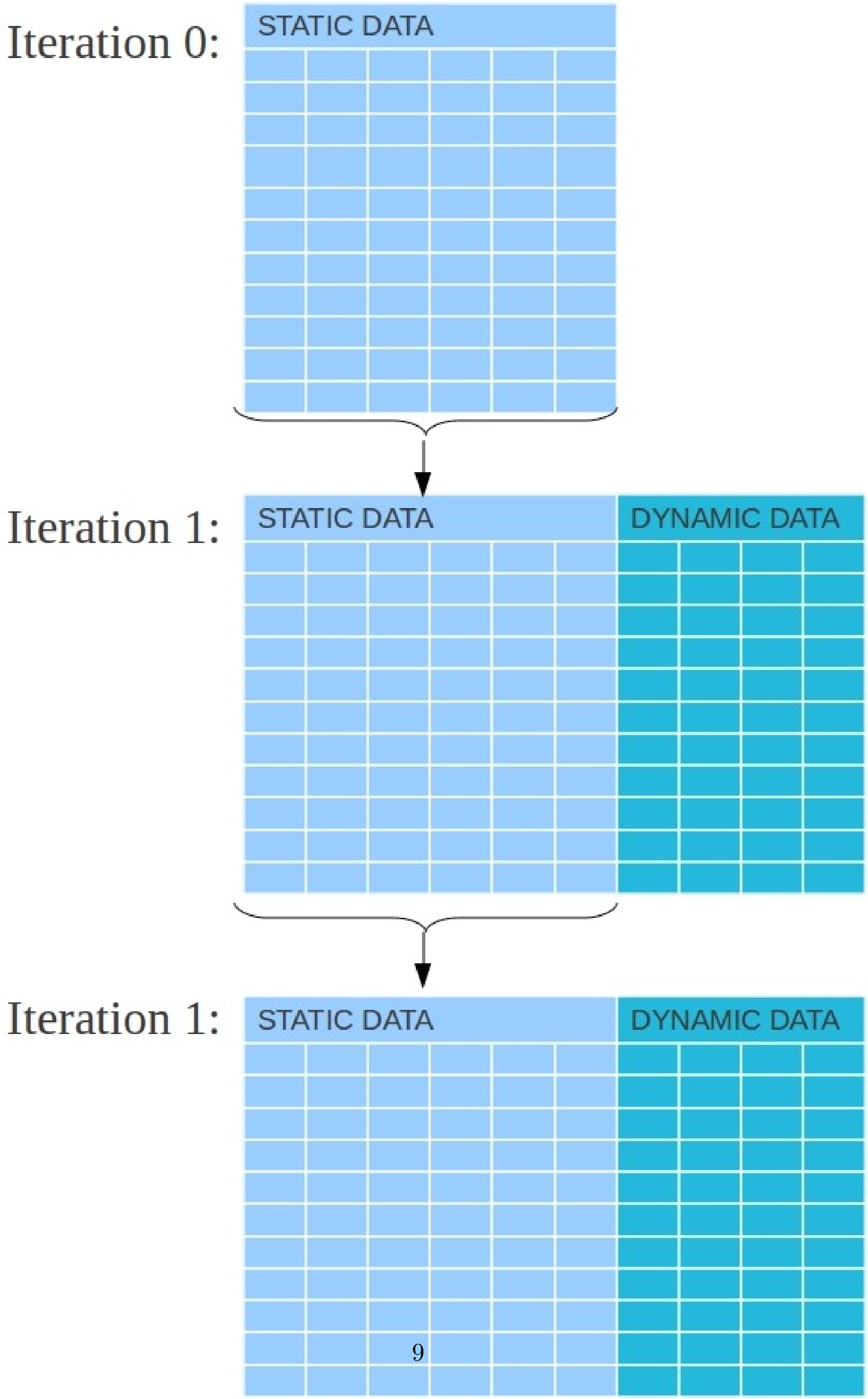
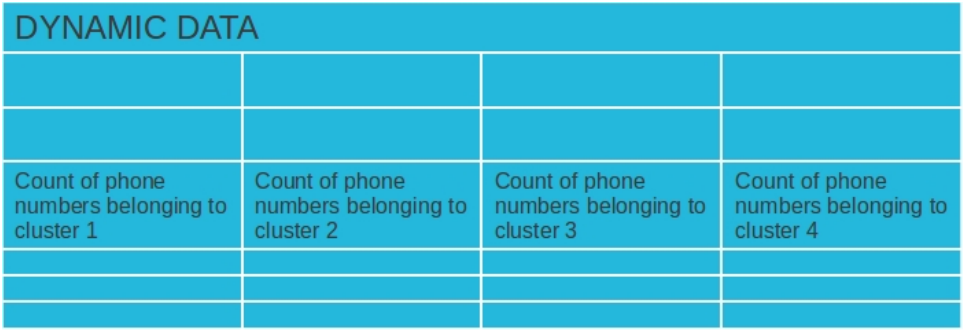


Figure 4: This figure shows how the dynamic data of each phone number is calculated for each iteration



DYNAMIC DATA			
Count of phone numbers belonging to cluster 1	Count of phone numbers belonging to cluster 2	Count of phone numbers belonging to cluster 3	Count of phone numbers belonging to cluster 4

7 Mathematical formulation

The above steps are mathematically representeded as follow: Let pn_j be the j^{th} phone number and let us represent pn_j by a static data part s_j and dynamic data part d_j, i.e. $pn_j = (s_j, d_j)$. If k is the number of clusters, then $d_j = (m_{j1}, m_{j2}, ..., m_{jk})$, where m_{jk} is the normalized count of phone numbers that pn_j is calling that falls in k^{th} cluster from the previous iteration. This formula is better represented in equation (2). The normalization for the $m_{j,k}$ is done by the sum of all the $m_{j,k}$'s. As the clustering scheme evolves d_i keeps on changing through every iteration. Now, let us represent a collection of all the d_j by D. Also, let us represent the total number of pn_j by jmax. Then, $D = (pn_1, pn_2, ..., pn_{jmax})'$. For the i^{th} iteration, let us represent the corresponding quantities by a superscript i.

Then,

$$s_j^i = s_j^{i-1} = s_j^0 \tag{1}$$

$$d_j^i = (m_{j1}^{i-1}, m_{j2}^{i-1}, ..., m_{jk}^{i-1}) \tag{2}$$

where, m_{jk}^i = normalized count of destination numbers belonging to k^{th} cluster called by pn_j in $(i-1)^{th}$ iteration

$$pn_j^i = (s_j^i, d_j^i) = (s_j^0, d_j^i) \tag{3}$$

$$D^i = (pn_1^i, pn_2^i, ..., pn_{jmax}^i)' \tag{4}$$

8 Results

The clustering results can be analysed in 2 parts - static and dynamic. The static results are the results derived from the matrix D^0 and the dynamic results are derived from the information added with each iteration. The dynamic results are derived from $\left[m_1^i|m_2^i|...|m_k^i\right]$.

Static results:

The static results correspond to the clustering analysis based on the static part of the data as described earlier. These results are derived from the centers and the sizes of the clusters. The centers of the clusters for the static variables are tabulated in the table 3

Cluster Number	numCalls	duration	weekend1weekday0	day1night0	out1in0	missed1
1	0.200	0.014	8.203	1.156	0.025	0
2	0.743	2.273	1.328	0.843	1.052	1.093
3	2.027	0.807	0.901	0.838	0.925	0.92
4	0.541	0.825	0.837	1.118	1.046	1.039

Cluster Number	SMS1	VoiceCall1	longDuration1
1	5.397	0.017	0.000
2	0.593	1.091	3.461
3	1.539	0.88	0.687
4	0.73	1.060	0.614

Table 3: The cluster centers corresponding to the static part of the data at the 2000^{th} iteration. Note that the values have been rounded off to 3 decimal places

The sizes of the clusters are tabulated in the table 4 below:

Cluster Number	Cluster Size
1	1
2	12
3	27
4	52

Table 4: The cluster sizes for the 2000^{th} iteration.

From tables 3 and 4, the following observations can be made:

1. The cluster sizes are very unevenly distributed with 1 of the clusters having only 1 phone number whereas the cluster with highest phone numbers has 52. This is not a particular case of only this iteration. It was observed that in each iteration, the trend of cluster sizes being 1, 12, 27 and 52 was repeated always. Even thought the combination of the clusters having these sizes change from iteration to iteration, the pattern of the above mentioned cluster sizes do not change. This indicates that the clustering

is fairly strong and the group of numbers that are in a particular cluster tend to move together into a different cluster in a different iteration.

2. **Profile of Cluster 1:**
These users make the least number of calls. This is evident since the cluster size is just 1. So the number of phone calls made or received by this cluster as a whole may be quite small. So, this profile of cluster actually reduces to the profile of a single phone number. This cluster makes the least average duration of calls, highest weekend calls, highest day time calls, lowest outgoing calls, 0 missed calls, higherst SMS calls, lowest voice calls and 0 long duration calls.

3. **Profile of Cluster 2:**
This cluster is made up of phone numbers which make moderate number of calls, highest average duration of calls, moderate weekend calls, low day time calls, high number of outgoing calls and missed calls and the lowest SMS calls. They also make the highest number of voice calls and long distance calls.

4. **Profile of Cluster 3:**
Even though this is not the largest cluster(with a cluster size of 27), this cluster has the highest number of calls. All other parameters are moderate for this cluster.

5. **Profile of Cluster 4:**
Even though this is the largest cluster with a cluster size of 52 which is nearly twice the size of the next highest cluster size, this cluster has very low number of calls which are mostly weekday calls and day time calls. This cluster has the least long duration calls and all other parameters are bordering on the low and moderate side.

The above inferences have been summarized in table 5

Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of calls:	Lowest	Low	Highest	Low
Average Duration:	Lowest	Highest	Moderate	Moderate
Weekend calls:	Highest	Moderate	Low	Lowest
Day-time calls:	High	Moderate	Moderate	High
Outgoing calls:	Lowest	High	Moderate	High
Missed calls:	0	High	Moderate	Moderate
SMS calls:	Highest	Lowest	High	Moderate
Voice calls:	Lowest	High	Low	High
Long duration calls:	0	Highest	Low	Low

Table 5: The summary of the sizes over 20 iterations

Dynamic results:

The dynamic results correspond to the clustering analysis based on the dynamic part of the data as described earlier. These results are derived from the centers and the sizes of the clusters. The results of this section are given below:

1. In the experiment performed for the validation of the above proposal, the dataset consists of 92 phone numbers and the optimum number of clusters is 4. The optimum number of clusters is obtained by clustering the raw data and observing certain parameters like betweenss, withinss and the Davies-Bouldin Index (DBI). As per the above description of the experiment, l=92 and k=4. The result of the clustering is an augmented table of the Raw data and the 4 normalized $m^i_{j,k}$. The normalization is such that the $m^i_{j,k}$ give a measure of the probability of the destination numbers belonging to a particular cluster (1, 2, 3 or 4) for each of our 92 phone numbers. The probabilities or the $m^i_{j,k}$ is plotted for all the 92 phone numbers. The colored lines indicate the probability of a phone number's calls to belong to a particular cluster number from among clusters 1, 2, 3 and 4. Here we show the plot for the 1^{st}, 20^{th} and the 2000^{th} clustering.

Figure 5: Clustering results after 1^{st} iteration

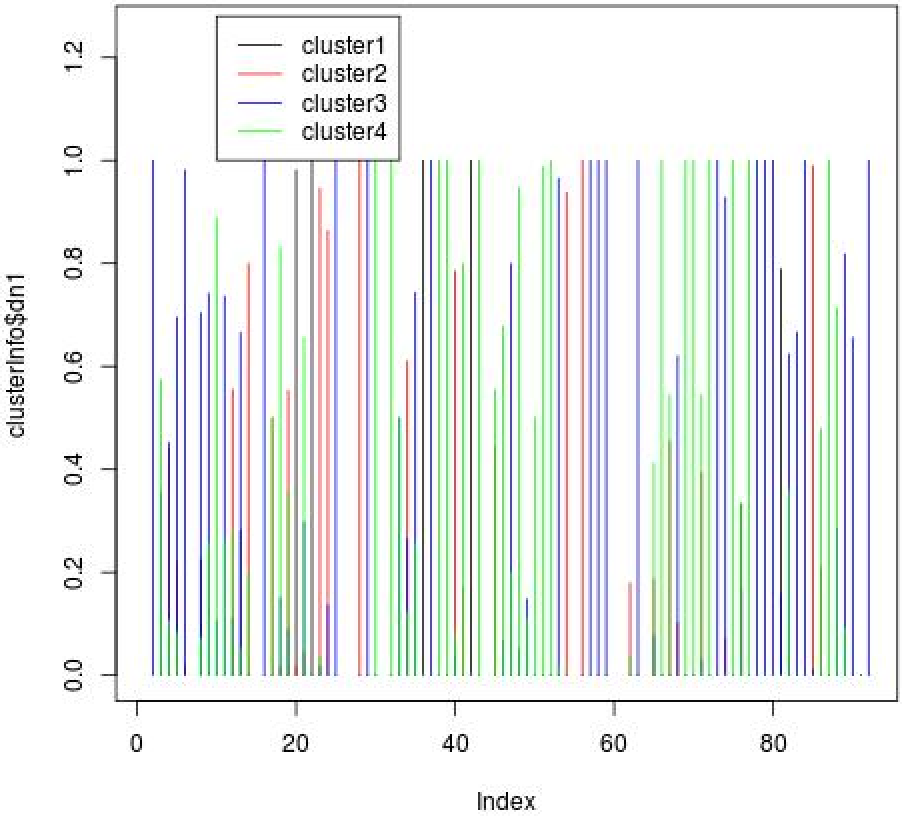


Figure 6: Clustering results after 20^{th} iteration

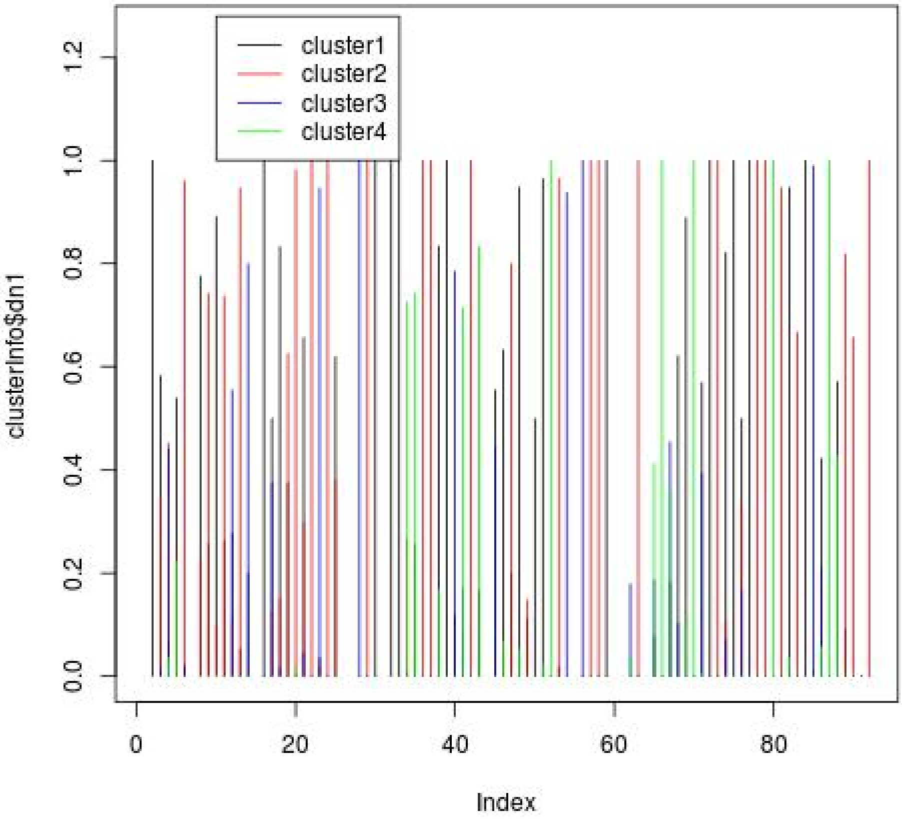
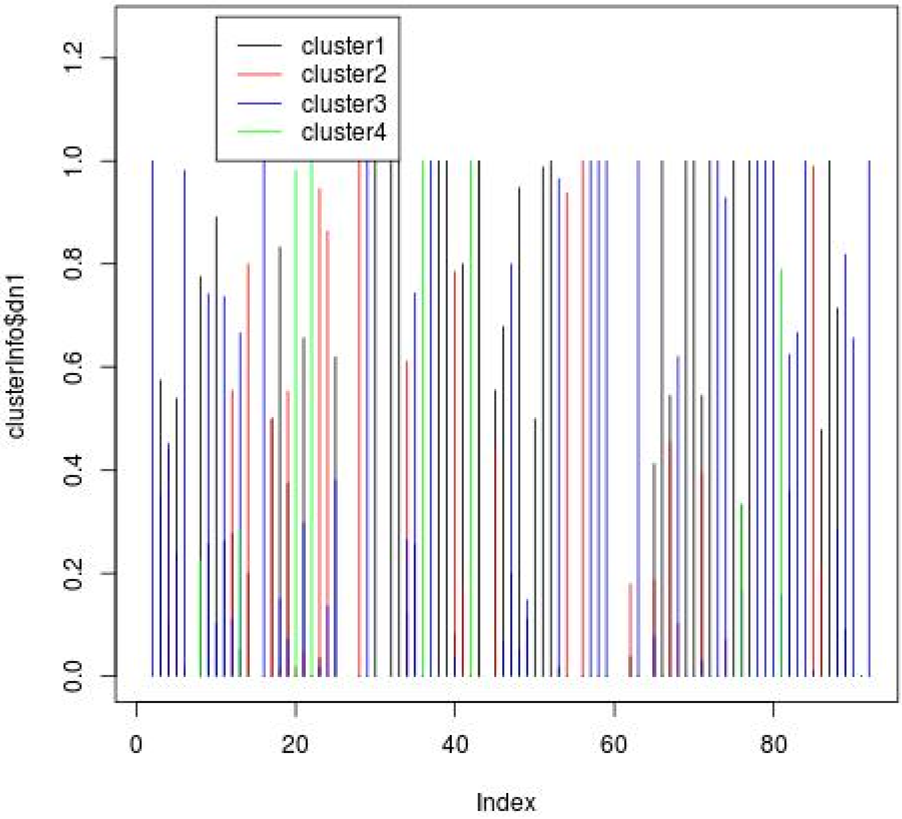


Figure 7: Clustering results after 2000^{th} iteration



The observations mentioned here are consistent with the other graphs too. The observations that we infer from the graphs are:

(a) Some of the numbers have high probability of their destination numbers to be in a particular cluster where as some of the numbers have a moderate probability.

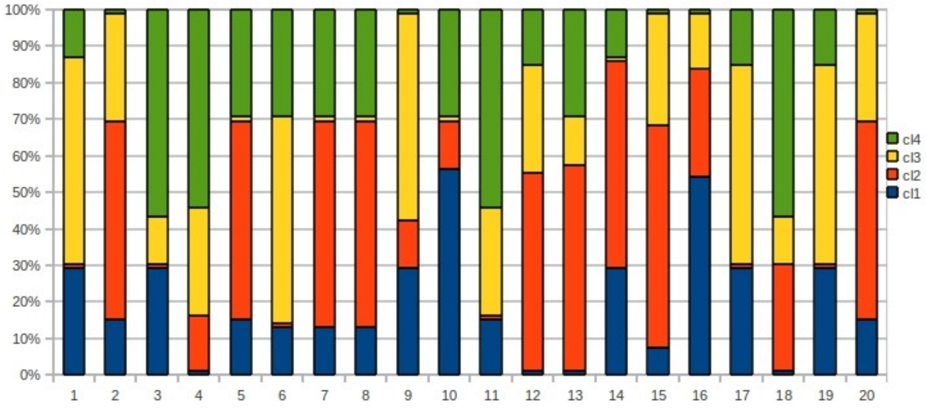
(b) The numbers with high probability of their destination numbers to belong to a particular cluster retain their high probability values through the iterations but the phone numbers with low probability values fluctuate a lot.

(c) Overall the basic structure of the graph does not change much over the iterations except for most of the probability values becoming higher marginally.

From the above observations, we can conclude that while certain phone users tend to concentrate their destination numbers to particular group of people(who fall within the same cluster because of their inherent calling behavior), others are more diversely networked. The phone numbers which have a closer social circle are the phone numbers which have a $m^i_{j,k}$ and the phone numbers with a more diverse social group have a low $m^i_{j,k}$.

2. A graph showing the relative sizes of the clusters for each iteration is given in figure 8.

Figure 8: Relative sizes of the clusters for 20 iterations



As we have mentioned earlier in the static results section, we see that the sizes are following a fairly standard pattern which we saw to be 1, 12, 27 and 52 even though they may be interchanging from cluster to cluster for each iteration.

3. The cluster centers for the dynamic part of the data can also be similarly analyzed. The cluster centers at the end of the 2000 iterations is given in the table 6 below:

Cluster Number	$m^i_{j,1}$	$m^i_{j,2}$	$m^i_{j,3}$	$m^i_{j,4}$
1	0.000	0.000	0.000	0.000
2	0.286	0.381	0.000	0.166
3	0.454	0.310	0.000	0.116
4	0.290	0.351	0.012	0.152

Table 6: The cluster centers corresponding to the dynamic part of the data at the 2000^{th} iteration. Note that the values have been rounded off to 3 decimal places

From table 6, it can be seen that the centers show some differentiation between the clusters even for a recursive clustering scheme. The results of the clustering as shown in table 6 is summarized below cluster-wise:

(a) **Profile of Cluster 1:**
The size of this cluster is 1 with the least number of calls as described in the static results section. In this section we get a further insight that this cluster does not make any phone calls to any of the numbers that we have used in the data set. If some phone calls were made, then, the cluster center of the cluser would not have been 0.

(b) **Profile of Cluster 2:**
The phone numbers of this cluster mostly make calls to other phone numbers in this cluster. This is indication of a closed social group. But these phone numbers do make some calls to other clusters except cluster 3.

(c) **Profile of Cluster 3:**
The phone numbers in this cluster make calls to other clusters but do not make calls to other phone numbers in their own cluster. The phone numbers in this cluster are mostly in contact with phone numbers from cluster 1.

(d) **Profile of Cluster 4:**
The phone numbers in this cluster make calls to all other clusters and they are the most socially diverse phone numbers.

9 Challenges faced

This section of the paper documents the challenges and pitfalls encountered in the course of the work so far. The challenges are enumerated below with details of their respective solutions.

(a) Making sense of the initial data - The initial data provided was a .mat file of 490MB which contained rich information about the call logs, bluetooth devices in the proximity of approximately 5 meters, cell tower IDs, application usage and phone status [5] for about 94 subjects. Most of the information provided by this dataset was not required for the purpose of the current project. So, the first challenge was to retrieve the relevant information from this large dataset. The README file [5] provided for the dataset was very useful to find out the necessary information. The relevant information was them exported as .csv files so that they can easily be manipulated and imported as databases.

(b) Text manipulation on the .csv files - From the previous step, the call logs of each phone was exported into a .csv file. Since there are about 106 phone numbers, there are also the same number of .csv files and these files need to be merged together before they can be imported into a SQL database. Each row of the .csv file corresponds to a phone log. And it is customary that each phone log contains information about the originating and the destination phone number. Since the data in the .mat file was organized in seperate tables for different originating phone numbers, the .csv files did not have information about the originating phone numbers. Hence, we need to do some text manipulation in order to introduce the information about the originating phone numbers. The steps followed are given below:

   i. Export the call logs of a phone into a .csv file whose filename reflects the originating phone number. e.g. the phone logs of phone number 4 should be exported into s4.csv.

   ii. Given that we have 106 files which are named as s-number.csv, we can go through each row of each file and a field at the beginning whose value is the number in the filename to which the log belongs. e.g. every log in s4.csv should have a column in the beginning with the value 4. This new column introduced at the beginning of each row can be used as the originating phone number. Due of the large number of rows to be changed, we run the following shell script in the same directory as that of the .csv files which does all the text manipulation as descibed above:

```
rename -v "s/\.csv/\.txt/g" *.csv
for ((i=1;i<107;i++))
nnndo
awk -v variable=${i}
'{printf variable;printf ",";print}'
s${i}.txt >> combine.txt
done
rename -v "s/\.txt/\.csv/g" *.txt
```

(c) Handling of date and time information - In the process of converting the .mat dataset into .csv files, certain fields contained information about the date and time of the calls placed. However, the date and time information was not in a recognizable form. These date and time information had to converted into string format by using the datestr function and then rewritten into the dateset before exporting into csv format. When importing the .csv file into a SQL table, the date and time information is first imported as a string field. The table is then updated with the appropriate SQL query to convert the string field into a date-time field. The SQL query for such a conversion is given below:

```
update combined_date set Date =
str_to_date(DateStr, '%d-%M-%Y %H:%i:%s')
```

After this conversion, the date and time information is available for logical manipulation as required.

(d) Importing the large database - The final comnbined .csv file is about 12.5MB. This can be easily imported using 'CSV with load' option which is much faster than other methods.

(e) Initially, normalization of the m_js was not done which gave very absurd and meaningless results. To realize that the reason for those absurd results were because of not normalizing the data was a long process.

(f) Keeping track of the iteration - There are several parameters in the new recursive algorithm and it is unknown as to what parameters are actually useful and what is not. So, to save the information as we recursively perform the algorithm and sort through several information to identify what information actually matters was difficult.

10 Future Work

This project has taken a new form of clustering and experimented on it for crisp clustering. The results are very promising with several new insights. Similar results are expected from fuzzy clustering, granular clustering etc. The future work will comprise of similar experimentation with these forms of clustering.

References

[1] BBC News, Over 5 billion mobile phone connections worldwide, 2010, doi: http://www.bbc.co.uk/news/10569081

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press:New York, 1981.

[3] D.L. Davies, D.W. Bouldin, "A cluster separation measure", IEEE Trans. Pattern Anal. Mach. Intelligence, vol. 1, 1979, pp. 224227.

[4] Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4:95-104.

[5] N. Eagle, "The Reality Mining Data README", 2010, doi:http://eprom.mit.edu/data/RealityMining_ReadMe.pdf

[6] N. Eagle, A. Pentland, D. Lazer, "Inferring Social Network Structure using Mobile Phone Data", Proc. of the National Academy of Sciences, vol. 106, no. 36, 2009, pp. 15274-15278.

[7] J.A. Hartigan, M.A. Wong, "Algorithm AS136: A K-Means Clustering Algorithm", Applied Statistics, vol. 28, 1979, pp. 100-108.

[8] H. Hohwald, E. Frias-Martinez, N. Oliver, "User Modeling for Telecommunication Applications: Experiences and Practical Implications", Proc. 18th Intl. Conference on User Modeling, Adaptation and Personalization (UMAP), 2010, doi:http://investigacion.tid.es/heath/images/mobile-user-models.pdf

[9] A. Joshi, R. Krishnapuram, "Robust Fuzzy Clustering Methods to Support Web Mining", Proc. the workshop on Data Mining and Knowledge Discovery, (SIGMOD '98), vol. 15, pp.1-8.

[10] M. Joshi, P. Lingras, C.R. Rao, "Correlating Fuzzy and Rough Clustering", Fundamenta Informaticae, in press.

[11] P. Lingras, "Unsupervised Rough Set Classification using GAs", Journal Of Intelligent Information Systems vol. 16, no. 3, pp. 215-228.

[12] P. Lingras, "Rough set clustering for Web mining", Proc. 2002 IEEE International Conference on Fuzzy Systems, 2002, pp. 200-205.

[13] P. Lingras, P. Bhalchandra, S. Mekewad, R. Rathod, S. Khamitkar, "Comparing Clustering Schemes at Two Levels of Granularity for Mobile Call Mining", Proc Rough Set and Knowledge Technologies (RSKT'11), in press.

[14] P. Lingras, C. West, "Interval Set Clustering of Web Users with Rough K-means", Journal of Intelligent Information Systems, vol. 23, no. 1, 2004, pp. 5-16

[15] P. Lingras, M. Hogo, M. Snorek, B. Leonard, "Clustering Supermarket Customers using Rough Set Based Kohonen Networks", Proc. Fourteenth International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence Series, 2871, Springer, 2003, pp. 169-173

[16] P. Lingras, M. Hogo, M. Snorek, "Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets", Web Intelligence and Agent Systems: An International Journal, vol. 2, no. 3, 2004, pp. 217-213

[17] J. MacQueen, "Some Methods fir Classification and Analysis of Multivariate Observations", Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297

[18] G. Peters, "Some Refinements of Rough k-Means", Pattern Recognition, vol. 39, no. 8, 2006, pp. 1481-1491.

[19] J.F. Peters, A. Skowron, Z. Suraj, W. Rzasa, M. Borkowski, "Clustering: A rough set approach to constructing information granules", Proc. 6th International Conference on Soft Computing and Distributed Processing, (SCDP 2002), 2002, pp. 57-61.

[20] Sharma SC, Werner A (1981) Improved method of grouping provincewide perma- nent traffic counters. Transportation Research Record 815:1318

[21] A. Skowron, J. Stepaniuk, "Information granules in distributed environment", in New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, N. Zhong, A. Skowron, S. Ohsuga, Eds. Lecture notes in Artificial Intelligence. Springer Verlag, Tokyo, Vol. 1711, 1999, pp. 357-365