IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A STUDY OF MOBILE PHONE POPULATION INDICATORS

[1]KEERTHANA B, [2]MINLA K.S

[1]Msc Scholar, [2]Assistant Professor

[1,2]Department of Computer Science

[1,2]St. Joseph's College (Autonomous), Irinjalakuda, Thrissur, India

ABSTRACT

Factual specialists advance and protect the creation and distribution of official measurements that serve the public great. One of their obligations is to screen the presence of people district by locale. Customarily this movement has been directed through censuses and overviews. These days innovations open additional opportunities like a nonstop detecting of the existences by utilizing the information related to versatile gadgets; example: The way of behaving of clients on doing calls. Different studies have demonstrated that data on mobile phone use is useful for analyzing problems in tourism, transportation planning, public administration etc. The methodology gives better populace assessment regarding cutting edge contrasting and genuine statistics information. The adaptability and flexibility that portrays the proposed structure empowers novel situations for the portrayal of individuals through information got from portable clients. In this paper, we survey a few advances made as of late in the investigation of cell phone datasets. We overview the commitments made such a long ways on the informal communities that can be built with such information, the investigation of individual versatility, geological dividing, metropolitan preparation, and help towards improvement as well as security and protection issues.

KEYWORDS

Clustering, Population indicators, Mobile Phones, Municipalities

INTRODUCTION

Mobile phones are one of the important components in our day to day life. Initially it was used for just communication but along the way with passage of time mobile phones now has a wide variety of functionalities. Currently, there are over six billion mobile phone users, with several hundred million expected to grow in the future. This enormous growth in the mobile phones users leads a way to explore how we can use the data on the mobile phones effectively. The data in the mobile phones can be used in many section such as education, transportation, medical etc.

In this paper we're discussing about a scalable and flexible solution for mobile phone based population indicators. The basic unit of mobile network infrastructure is a cell with its own cellular base station. The base station is a fixed transceiver that is the main communication for one or more wireless mobile client devices. Whenever a cell phone user moves from one base station to other, the network automatically switches to the other respective base station. By using this information we can assume one person's work place and residence. For this we need to have classify the individuals sharing common behavior into groups. For this the algorithm used was k-means. But k-means clustering has certain challenges.

1) It has to be run for a certain amount of iteration for getting the result.
2) Unable to handle high volume data.
3) Displays memory error in large databases.

Here we are introducing a base version of Muchness [1], a framework to overcome these challenges. Using Muchness we are able to estimate the number of residents, passengers and visitors in a give region by using mobile phone data. We defines a similarity metric which is able to capture the similarities in the calling behavior of users as well as the number of calls completed. We will study the movements of individuals between home and work and to other places. Based on the result we estimate the number of residents, passengers and visitors which will allow us to identify the similarities between the individuals and thereby we introduce individual profiles.

RELATED WORKS

The mobile phones data has been used in many areas. It can be used to monitor traffic in cities and to analyze the movements of tourists. In a case study in Rome [2] on real-time urban monitoring using cell phones. They uses a Localizing and Handling Network Event Systems (LocHNESs). It was developed by Telecom Italia. By studying the positioning of buses and taxis they provide information about the mobility. There were other works for instance the winner of the Nokia Mobile Data Challenge [3]. Using data collected from the smart phones of about 8 users, they explored the characteristics of the user's mobility traces. Then they developed three families of predictors, including tailored models and generic algorithms, to predict, based on instantaneous information only, the next place a user will visit. De Jonge [4] had a different approach. They looked at a dataset of mobile phone call activity to see if it might be used in official statistics. These dataset provides longitudinal, geospatial indicators that relate to economic and cultural activity. They looked at the mobility of mobile phone users by using logged call-events and compare the results of this analysis to official mobility statistics. Another work from Terada [5], The population Estimation Technology for Mobile Spatial Statistics gives explanation of the approach to estimation and the estimation methods used for MSS, which is used to make estimations of population using the operations data from a mobile terminal network. But this resulted in some errors. Concluding an individual is resident by only checking the presence in a cell resulted in some errors. From the ideas of De jonge Deville [6] used datasets of more than 1 billion call records from Portugal and France. They proposed a framework called MP. They showed how maps of human population changes can be produced over multiple timescales while preserving the anonymity of MP users. With similar data being collected every day by MP network providers across the world the population density is estimated as a function of night time calls occurring in an area. However a simple rule based approach more useful information about the calling behavior of the users cannot be collected. To overcome these limitations, Furletti [7] investigated to what extend such mobile data could be a support in producing reliable and timely estimates of inter-city mobility. By using the data they build individual profiles that assumes the places of residence and that of work (or study). The distance between a person’s home and their place of work (or education) is used as a proxy for their lack of intercity mobility (we define him a static resident). By analyzing this it is possible to identify whether an individual is resident, commutator or visitor. Sociometer [8] is a framework aimed at classifying mobile phone users into behavioral categories by means of their call habits. By using k-means algorithm, they focused to aggregate users by learning different behaviors, and returns annotated profiles.

MUCHNESS: A FRAMEWORK

Muchness is a framework to derive statistics about population by grouping them by means of calling behavior. Then we analyze the data and classify the individuals as resident, commuter or visitor. This clustering algorithm provides scalability and it is designed for a distributed environment. Our algorithm doesn’t need to the number of clusters in advance. Muchness is an advanced version which overcomes the drawbacks of k-NN based algorithm [9] and Sociometer [8]. In the next sessions we will describe how data is collected and aggregated and the metrics used to classify individuals having similar behaviors.

A. Data Description

For billing purposes telecom operators collect customer data. Every call record consists of a tuple having the anonymous identifier of the user, the call timestamps and the cell id contains an overview of how data are created, collected and aggregated. We need to identify whether an individual is resident, commuter or visitor. For this from a telecom operator in Italy, we have obtained anonymous data of calls made in Tuscany (Italy) recorded during the period between February and March 2014. We processed about 2.6 million records representing calls. We had about 800k individuals call records. By analyzing the data we create a user calling profile (ICP) which contains the user id, data, time, cell id etc. For every person there will be an individual calling profile (ICP). It represents the calling behavior of a person. If a person performed calls to different municipalities they may have multiple ICPs. These are used to determine whether an individual is a resident, traveler or visitor to the municipality. Each ICP is a 30-dimensional matrix in which each position represents a specific time interval of the day (morning, afternoon, night), between weekdays and weekends for a total of 5 analyzed weeks.

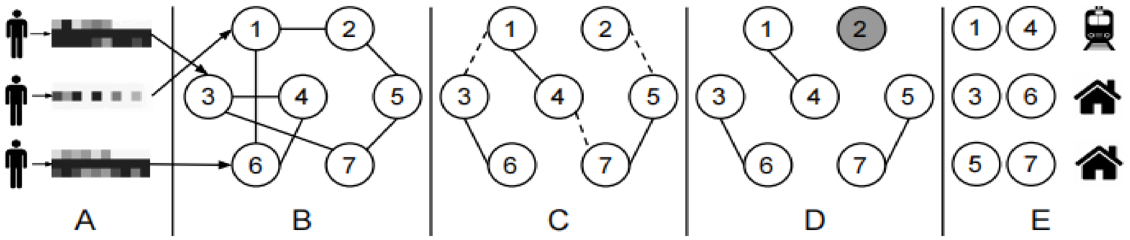


Fig. 1: Muchness analytical process. A: assigning ICP for each individual . B: each ICP becomes a node. C: we search for similar nodes and at the end we cut off low similarity edges. D: Connected components are identified and define the outliers (node 2). E: for each cluster we define an exemplar to classify as Resident, Commuter or Visitor.

B. The clustering algorithm

The clustering algorithm we use is a k-NN based text clustering algorithm [9]. In the following of this section we provide a short description of the primary functions characterizing such work to help understanding how to choose the ideal parameter values, a way to introduce specific metrics and help understanding the improvements introduced.

1. The analytical process: In figure 1, for each mobile user we build an ICP (see column A). Then, we generate a graph of ICPs. At start we randomly link each node to few other nodes (see column B). Then, the algorithm iterates, starting from the initial graph, adjusting the neighborhood of each node with most similar nodes. The edges connecting nodes with similarity less than a predetermined threshold parameter are then trimmed (see column C). The resulting clusters are the connected components [10] (column D). As a final step, exemplars are generated (column E). where clusters are classified as Commuter, Visitor or Resident based on automatic classification.
2. Parameter choice: There are two specific parameters: k and ε. k represents the number of neighbors for each node in the graph. ε is threshold parameter that drive the edge pruning process to avoid that very different nodes would fall in the same cluster. k represents both the quality and the execution time of the clustering. The clustering technique stars with a randomly connected graph and connects each node to its k most comparable nodes based on a similarity measure. The similarity measure can be arbitrary. Here we are keeping the connections between the high similarity nodes and not keeping the connection between the low similarity node.
3. Adapting the algorithm for ICPs analysis: To improve the algorithm to suite the ICPs data we define two processes. One is injecting an arbitrary similarity matrix. Here we define specific similarity metrics that are able to exploit the similarity between the ICPs data. Other one is the exemplar definition, due to the large size of the dataset it is necessary to define an exemplar for each cluster. The exemplar is the first entry point to analyze a cluster by a manual investigation.

C. Metrics to capture ICPs similarities

Each ICP is a 30 dimensional array representing the calling behavior of an individual. There are certain metrics to process these data. We need to introduce a metrics to capture the similarities between individuals. Next, we present our metrics to improve the quality of the results obtained by the clustering (Table 1).

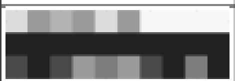
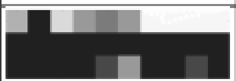
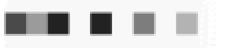
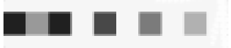
			EUC	JAC	EUC+JAC
Residents			0.5	1	0.8
Commuters			0.78	1	0.91

Table 1: ICPs extracted by experts with similar characteristics. A comparison of similarity values using: EUC, JAC and EUC+JAC

1. How to capture shapes similarity: Here we use Jaccard similarity (JAC), to use JAC we modify each array in a Boolean array where we set the value 1 in position i if in position i the data has a value greater than 0. However, the JAC doesn't take into account about the weights in the array. Therefore we combine the two similarities, the EUC and the JAC where;

   $$EUC+JAC(a,b) = \alpha EUC(a,b) + (1-\alpha)JAC(a,b) \quad (1)$$

2. Comparing the metrics, an example: In Table 1 the first two columns is the ICPs selected and the last three is the similarity in the values using different metrics. The ICPs have a similar behavior resulting in similar shapes. For example, take the first row in Table 1 and consider the two residents. Although some positions have different values, note the color darkness representing the value on a single position of the array, they have an equal shape representing the same calling behavior. It is not possible to determine whether the two ICPs are similar (only 0.5 similarity), but the JAC (giving value 1) suggests that they are the same shape. We can take advantage of both measurements with our EUC+JAC and achieve a high similarity of 0.8.

EXPERIMENTAL EVALUATION

All the experiments have been conducted on a cluster running Ubuntu Linux 12.04 consisting of 5 nodes (1 master and 4 slaves), each equipped with 128 Gigabytes of RAM and with two 16 cores CPU, which is connected using an ethernet network. The approach was implemented using Appache Spark [11], the source code we used for conducting our experiments is publicly available on GitHub.

HOW TO CONFIGURE MUCHNESS

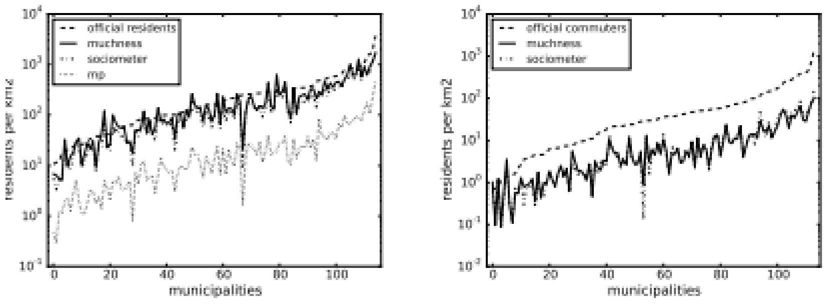
Our approach needs proper values for k and ε from the input from the users. We are introducing a new strategy to make this selection easier. Following that statement, we tested our approach distribution to k values among the vary [5,10] getting terribly similar results each in terms of internal cluster evaluation and in terms of residents and commuters identified. Another parameter we use here is the parameter ε is intended to be used for conducting a preprocessing phase in which all the edges below such value are pruned before starting with the cluster identification process.

INTERNAL CLUSTERING EVALUATION

Describes how we evaluate the internal clusters.

- Compactness: It is a metric that determines how near elements in a cluster are to one another. The average pair wise similarity among items in each cluster is used to calculate compactness. The higher the value, the better.
- Cluster Separation: This metric assesses how effectively clusters are separated from one another. Computing is used to achieve separation the average degree of similarity between items in various groupings. Lower numbers are preferable.

COMPARING WITH OFFICIAL STATISTICS BUREAU



(a) Residents (b) Commuters

Fig 2. Comparing with Official Statistic Bureau

	Residents $\times km^2$			
	<50	50 - 100	100 - 150	>150
MP	93%	91%	92%	94%
Sociometer	39%	39%	49%	52%
Muchness	**24%**	**29%**	**42%**	**47%**
	Commuters $\times km^2$			
Sociometer	**83%**	84%	86%	89%
Muchness	84%	**83%**	**81%**	**87%**

TABLE II : Median estimation errors comparison

	Residents	Commuters	Visitors
MP	74 021	N/A	N/A
Sociometer	405 845	21 549	2 224 575
Muchness (EUC)	137 121	7 148	2 231 323
Muchness (JAC)	407 020	12 394	2 175 692
Muchness (EUC+JAC)	432 047	15 187	2 037 022

TABLE III : Number of Residents , Commuters and Visitors

In this section we compare our results against MP [6] and Sociometer [8]. To study the performances of Muchness with respect to alternative existing approaches, we compared against the following competitors the first one is Sociometer, which is the primary competitor, it is the most similar to Muchness; both the approaches are based on clustering and designed for the same case study MP targets the same problem, it relies on rules rather than clustering, such as the calling hours to determine if a person is a resident. dbscan, we tried to conduct our experiments with an implementation of MR-dbscan, unfortunately we have not been unable to cluster more than 10% of the dataset due to memory errors .The results are compared to official census. Statistics from the Italian National Statistical Institute, This data includes the number of residents and commutators belonging to the 115 municipalities we studied. First, let's look at the total number of residents, travelers and visitors. Table III shows the results. MP provides a significantly lower population count than Muchness, Sociometer.

Next, we compare these results to the actual census provided by ISTAT. Figure II shows the number of identified residents. The y-axis shows the estimated population density, while the x-axis shows the municipalities ordered from lowest to highest population density. All methods have peak values in the same municipalities. Accordingly, despite the different approaches (MP defines rules, Sociometer determines clustering, and ours does clustering), they all recognize similar behavior. It is evident that MP always underestimates density with a larger error than Sociometer and Muchness. We divided the error of the estimates into 4 areas with different population densities. Table II shows the mean error of the estimates. Again, Mp provides the estimated value affected by the largest error. Muchness and Sociometer provide similar results for higher density municipalities where the volume of available data is large and clustering can be based on a variety of information. In contrast, Muchness provides a 10% lower error rate than Sociometer for less dense communities, particularly those with a density less than 50 or in the range of 50 to 100 km^2. Then we are compare the results against real census data. Both methods produce nearly identical outcomes.

CONCLUSION

The telecom operators has a lot of information such as handling the localization of mobile phones, optimizing the capacity of a site, handling billing information, is stored by the mobile phone company. So mobile phone companies record data that are very closely associated with behavior of people. Nowadays the mapping of population is done by means of surveys and logistic of censuses. These methods have lot of challenges and issues. Here we are introducing a framework to estimate the population using the mobile phone data. With respect to the existing works, we uses clustering algorithm to capture similarities between individual call profiles. Evaluation results show that the estimated population is well related with the census population. The findings in this study will help us to understand population dynamics precisely.

REFERENCES

1. Lulli, A., et al.: Improving population estimation from mobile calls: a clustering approach. In: 2016 IEEE Symposium on Computers and Communication (ISCC). IEEE (2016)
2. Calabrese, F., et al.: Real-time urban monitoring using cell phones: a case study in rome. IEEE Trans. Intell. Transp. Syst. 12(1), 141– 151 (2011)
3. Etter, V., et al.: Where to go from here? Mobility prediction from instantaneous information. Pervasive Mob. Compute. 9(6), 784–797 (2013)
4. De Jonge, E., van Pelt, M., Roos, M.: Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. In: Paper for the Federal Committee on Statistical Methodology research conference, Washington, USA (20
5. Terada, M., Nagata, T., Kobayashi, M.: Population estimation technology for mobile spatial statistics. NTT DOCOMO Techn. J. 14, 10–15 (2013)
6. Deville, P., et al.: Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111(45), 15888–15893 (2014)
7. Furletti, B., et al.: Use of mobile phone data to estimate mobility flows. measuring urban population and inter-city mobility using big data in an integrated approach. In: Proceedings of the 47th Meeting of the Italian Statistical Society (2014)
8. Gabrielli, L., et al.: City users' classification with mobile phone data. In: 2015 IEEE International Conference on Big Data, pp. 1007–1012 (2015)
9. A. Lulli et al., "Scalable k-nn based text clustering," in Big Data, 2015 IEEE International Conference on. IEEE, 2015, pp. 958–963
10. A. Lulli et al., "Cracker: Crumbling large graphs into connected components," in Proc. of IEEE ISCC, 20
11. "Apache spark," https://spark.apache.org