Identification of Users' Geographic Map

Cláudia B. Rodrigues
CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
cbarodrigues@student.dei.uc.pt

Marco A. Veloso
Polytechnic of Coimbra, ESTGOH
CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
mveloso@dei.uc.pt

Ana O. Alves
Polytechnic of Coimbra, ISEC
CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
ana@dei.uc.pt

Carlos L. Bento
CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
bento@dei.uc.pt

Abstract

The capacity of improving urban spaces strongly depends on having data that supports decision-making and provides valuable information for planning. In this paper, we aim to study anonymized Call Detail Records (CDRs) of clients in Coimbra (Portugal) and adopt clustering algorithms to obtain their geographic map by identifying their most visited places, at an antenna level.

We propose a new methodology to identify Home, Second Home, and Workplace, assuming that clients have different routines. We apply clustering to segment customers and profile them according to their daily CDRs. Based on identified profiles, sleep and work hours are extracted and a density-based algorithm is applied to recognize their places. Ground-truth is used to validate and evaluate the model on the inference of daily locations.

2012 ACM Subject Classification Geographic information systems

Keywords and phrases Call Detail Records, Clustering, Data Analysis, Geo-Data, Geo-Profile, Meaningful Places, Sleeping Period

1 Introduction

Patterns in human routines tend to be quite predictable, presenting a high degree of temporal and spatial regularity [6] [9]. The study of trajectories, generally leads to the creation of models to identify individual's mobility patterns that, most of the time, can faithfully reproduce their movements. Ultimately, this is a task that relies on geospatial data, such as CDRs. This type of data is used for location analytics to characterize various aspects of human mobility, namely, improving the public transportation systems or deploying new services or infrastructures [11].

In this work we propose a model for individual geo-profiling, using the users' records on the mobile network. The geo-profile consists of identifying meaningful places (places that belong to the user's routine), such as home, second home, and workplaces, without assuming that all users have the same routine profile.

The paper is organized as follows: in section 2 a brief review of the related work. In section 3 we discuss the used dataset. In section 4 we introduce the methodology to identify profiles and geo-profiles of users. Section 5 presents the experimental work and results. In section 6 we highlight the main achievements with this work.

2 Related Work

Mobile phones are worldwide available and ubiquitous, and the study of human movements through the exploitation of mobile phone data has been an active area of research. CDRs are provided by mobile operators, without requiring the users' participation, with the advantage of being available for significant groups of the population and once anonymized are not intrusive. Its analysis provides knowledge on the user's sent and received calls and text messages, as well as about the antenna that received/transmitted the communication.

Although CDRs present significant challenges as a source of location information, such as temporal irregularity and spatial sparseness, many researchers and institutions are aware of their potential. Thus, they ended up showing that it can reflect human mobility and significant places, considering this data representative and using it to achieve goals, such as the identification of meaningful places [7] [11].

Therefore, some authors used CDRs to identify home and workplaces. Some of them rely on the user's common behavior, using *a priori* assumptions (e.g.: criteria with temporal constraints) to determine important places [7] [14] [13].

Although being common the application of rules/criteria to infer home and work locations, their use implies that subscribers with different routines are treated the same way as the subscribers that have common routines. To identify important places for different types of users, with distinct habits, some researchers did not make any *prior* assumptions on the behavior of the users [10] [3] [12].

Previous research showed that it is possible to identify meaningful places of users with different behavior through the analysis of CDRs. Thus, we attempt to improve the profile identification method, to better understand the users' routines.

3 Dataset Description

In this paper, we analyze the anonymous CDRs of 36 000 users from July to October of 2020. These records include information on how, when, and where people daily communicate. Each CDR is designated as an event and in our dataset, contains the following information: ID of the caller, ID of the antenna, geographic location of the antenna, coverage radius of the antenna (in meters), and information of when the event took place.

The data was analyzed, filtered, and prepared, eliminating irrelevant columns or deriving new necessary attributes. Following Ahas et al. [1] work, it is assumed that clients with events in their most visited cell on fewer than seven days a month, are not suitable and their geo-profile is an unsuccessful task to perform. So, they were excluded from the final dataset.

4 Methodology

After the data preparation (exploratory analysis and filtering), the dataset contains only users that have the relevant events to be geo-profiled. The next steps consist of using a clustering algorithm, K-Means [4], to segment the users and characterize the resultant groups, to infer periods that are assumed to be part of their sleeping period and working hours. This algorithm has the advantage of generating groups of users of different sizes. Besides, an asset of its usage is that we know, *a priori*, the number of groups that will be generated.

Then, a density-based algorithm, VDBSCAN [8], is used to infer meaningful places for each user based on his/her routine features. This algorithm is an extension of DBSCAN, which identifies places where the user has a significant number of events [5]. VDBSCAN is also able of dealing with outliers and datasets with varying densities. The selection of the algorithm was based on works that used DBSCAN to identify home and work locations [14] [3] and the necessity of dealing with datasets with varying densities of the antennas.

The final step is the validation and evaluation of the obtained results. To perform this, we use as ground-truth data the centroid of the postal-code (ZIP code), from a percentage of the clients within the geographic area of analysis.

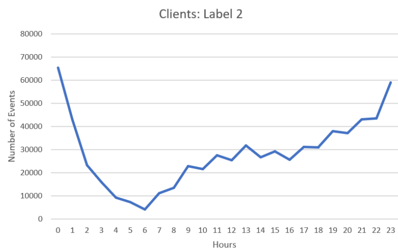
5 Experimental Results

5.1 Customer Segmentation

We started by studying the wake cycle of the clients based on their activity on the network (CDRs) and realized that most of them varied their sleep time patterns between workdays (Monday to Friday) and weekends (Saturday and Sunday). We concluded that the best scenario to group subscribers was based on the events made during the night (12AM-4AM), the morning (5AM-10AM), and the afternoon (1PM-5PM) on workdays. After analyzing the mean number of events that each user made throughout these slots, the elbow/knee method [2] returned an optimal k=3 for K-Means.

We identified three groups of users with different routines according to the number of CDRs that they have throughout the workdays. Based on the analysis of the activity and the Portuguese work code, we determined slots of sleeping periods and work hours for each group of users. For Group 1, in figure 1(a), constituted by "*day workers*", the sleeping period was determined between 2AM and 5AM, and work hours from 9AM to 1PM. The users in Group 2 (figure 1(b)), were identified as "*night workers*", and their sleeping period from 5AM to 9AM, and working hours from 12AM to 4AM. The routines of the users in Group 3 are not directly visible, so based on the common sleeping period in Portugal, it was assumed to be from 12AM to 4AM and the working hours from 1PM to 5PM [7].

(a) Group 1



(b) Group 2

Figure 1 Customer segmentation

5.2 VDBSCAN: Home, Second Home, and Workplace

VDBSCAN was selected because the density of antennas is higher in urban areas than in rural areas. A characteristic of this algorithm is that the *eps* parameter is automatically calculated according to the density of the dataset, however, the *MinPoints* parameter has to be manually defined [8].

For the home location, since we are analyzing a sleeping period, it is necessary only one event along that period to determine the home location. If there are no events throughout this period, the events on the two hours before and after, are collected. In cases that even after this, there are no events registered, it is declared impossible to identify home locations.

To identify the work location, it is necessary that the user has at least five events during the five hours defined as work hours.

The identification of the second home does not rely on the segmentation. To determine this place, it is assumed that second home locations are where the user has more than five events from 7PM to 7AM on workdays and weekends from July to August, which in Portugal is the common vacation season, and on weekends in September and October.

5.3 Results

The accuracy evaluation was performed for home locations because there is no profile information about users' second home and workplace. From the 4 600 users on the ground-truth dataset, we were able of identifying home locations for 3 838 clients (83.4%).

The method to evaluate the results was used by Mamei et al. [10] and consists of applying the coverage radius of the antennas to declare if the place was accurately found or not. Considering the different densities of the antennas in urban and rural spaces, we distinguish these areas to achieve the mean coverage radius of the antennas in each one. With our data, the mean coverage radius in urban areas is 1 900m and 2 820m in rural areas.

The results are evaluated in a way that if in an urban or rural area, the distance between the antenna found as home and the real home location, is lower than the coverage radius of the respective area, the home location is considered achieved with success.

The results of the evaluation are presented in table 1. After performing multiple experiments, we considered that, for the evaluation, distances between the two places higher than 20km, should be treated as annotation errors. We assumed that users that presented this scenario, were living in the home identified by the model, however, their billing address was registered far away from that place.

Table 1 Scenarios/Situations identified

Situation	Note	Description
Type 0	**Annotation Errors**	Outliers (20% of the 3 892 homes identified)
Type 1 (66%)	**Success**	The antenna identified is near the real home location (Distance <1 900/2 820m)
Type 2 (15%)		The antenna is the nearest from the real home, but the distance is bigger than the established
		The antenna identified to represent the home location is not the nearest but is close
Type 3 (19%)	**Unsuccess**	The home location is not properly determined

6 Conclusions

This paper proposes a method for identifying significant places. As illustrated in table 1, we reached an accuracy of 66% on the identification of home locations (Type 1). We also found a situation that we considered a success scenario with less degree of certainty (Type 2), which we assumed to be caused by the large coverage area of the antennas.

We used a technique by Lumpsum et al. [12] to infer sleeping periods and compare our results: for each user, we identified an hour with less activity to determine the sleeping hour and encounter the sleeping period based on that. These authors obtained an accuracy of

69.02%. However, with our data and their technique we obtained an accuracy of only 60%. The achievement of these results led us to profile the users at a group level, using K-Means.

The results were validated and evaluated with real and updated data and using a method that was used by other authors [10]. Our future work will focus on improving the segmentation and the analysis of routines, and continue the process of validation and evaluation of work and second home results, which will be performed using data from a survey.

References

1. Rein Ahas, Siiri Silm, Olle Järv, and Erki Saluveer. Using mobile positioning data to model locations meaningful to users of mobile phones. In *Journal of Urban Technology*, pages 3–27, 2010. `doi:10.1080/10630731003597306`.
2. BASILB2S. In-depth intuition of k-means clustering algorithm in machine learning. URL: `https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/`.
3. Oliver Burkhard, Rein Ahas, Erki Saluveer, and Robert Weibel. Extracting regular mobility patterns from sparse cdr data without a priori assumptions. In *Computer Science, Journal of Location Based Services*, 2017.
4. Imad Dabbura. K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. URL: `https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a`.
5. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231, 1996.
6. Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. In *Nature*, volume 453, page 779–782, 2008.
7. Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, and James Rowland. Identifying important places in people's lives from cellular network data. In *Pervasive Computing - 9th Intl. Conf., Pervasive 2011, San Francisco, CA, USA, June 12-15, 2011*, 2011. `doi:10.1007/978-3-642-21726-5_9`.
8. Peng Liu and Dong Zhou andNaijun Wu. Vdbscan: Varied density based spatial clustering of applications with noise. In *Intl. Conf. on Services Systems and Services Management, ICSSSM*, 2007.
9. Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. 2013.
10. Marco Mamei, Massimo Colonna, and Marco Galassi. Automatic identification of relevant places from cellular network data. In *Pervasive and Mobile Computing*, 2015.
11. Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? In *ACM SIGMOBILE Mobile Computing and Communications Review*, pages 33–44, 2012. `doi:10.1145/2412096.2412101`.
12. Lumpsum Tongsinoot and Veera Muangsin. Exploring home and work locations in a city from mobile phone data. In *IEEE 19th Intl. Conf. on High Performance Computing and Communications, IEEE 15th Intl. Conf. on Smart City, IEEE 3rd Intl. Conf. on Data Science and Systems (HPCC/SmartCity/DSS)*, 2017.
13. Maarten Vanhoff, Fernando Reis, Thomas Ploetzand, and Zbigniew Smoreda. Detecting home locations from cdr data: introducing spatial uncertainty to the state-of-the-art. In *Computer Science*, 2018.
14. Peiyu Yang, Xuejin Wan, Tongyu Zhu, and Xuejiao Wang. Identifying significant places using multi-day call detail records. In *IEEE 26th Intl. Conf. on Tools with Artificial Intelligence*, 2014. `doi:10.1109/ICTAI.2014.61`.