Pervasive and Mobile Computing 31 (2016) 147–158

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Automatic identification of relevant places from cellular network data

Marco Mamei [a,*], Massimo Colonna [b], Marco Galassi [a]

[a] DISMI, University of Modena and Reggio Emilia, Italy
[b] Engineering & Tilab, Telecom Italia, Italy

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 8 September 2015
Accepted 7 January 2016
Available online 29 January 2016

Keywords:
Mobility patterns
CDR data
Place identification

ABSTRACT

We present a methodology to automatically identify users' relevant places from cellular network data.[1] In this work we used anonymized Call Detail Record (CDR) comprising information on where and when users access the cellular network. The key idea is to effectively cluster CDRs together and to weigh clusters to determine those associated to frequented places. The approach can identify users' home and work locations as well as other places (e.g., associated to leisure and night life). We evaluated our approach threefold: (i) on the basis of groundtruth information coming from a fraction of users whose relevant places were known, (ii) by comparing the resulting number of inhabitants of a given city with the number of inhabitants as extracted by the national census. (iii) Via stability analysis to verify the consistency of the extracted results across multiple time periods. Results show the effectiveness of our approach with an average 90% precision and recall.

1. Introduction

The widespread diffusion of mobile phones and cellular networks provides a practical way to collect location-based information from a large user population. The analysis of such collected data is a fundamental asset in the development of several services and applications. In particular, we propose an innovative approach to automatically identify the places that people routinely frequent (e.g., home, work place, favorite nightlife locations) from the analysis of anonymized positioning data from a cellular network (i.e., CDR—Call Detail Record). The knowledge of such relevant places finds important applications in mobile services, marketing, traffic forecasting, urban planning and management services [1–4] (see Section 2 for more details).

The basic approach to extract places from CDR data, adopted by almost all the state of the art, is illustrated in Fig. 1. (1) CDRs of each users are collected. (2) CDRs are clustered in well specified spatial regions. (3) Clusters are then weighed on the basis of some factors (e.g., number of days in which the user visits the cluster). (4) Clusters with a weight greater than a certain threshold are associated to relevant places. Our proposal improves the state of the art in all the three phases: clustering, weighing, thresholding.

1. **Clustering.** Some of the approaches at the state of the art try to identify relevant places on the basis of the network cell from where most traffic is generated [5–7] (i.e., they cluster CDRs by the cell in which they are generated). This approach however tends to be inefficient in that, in areas covered by multiple cells, a cell phone will split CDRs among different

* Corresponding author. Tel.: +39 0522522233; fax: +39 0522522312.
E-mail addresses: marco.mamei@unimore.it (M. Mamei), massimo.colonna@telecomitalia.it (M. Colonna), marco.galassi@unimore.it (M. Galassi).

[1] The approach described in this paper is the subject of Patent Application PCT/EP2014/058003 filed by Telecom Italia on April 2014.

http://dx.doi.org/10.1016/j.pmcj.2016.01.009
1574-1192/

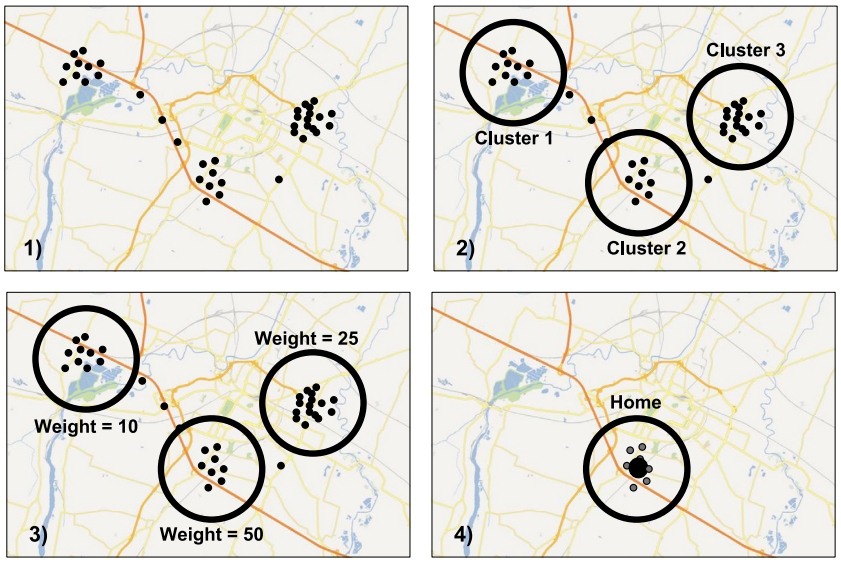

Fig. 1. General methodology to identify frequented places. (1) CDRs for each user are selected. (2) A clustering algorithm groups CDRs into well-defined spatial regions. (3) A weighing mechanism gives each cluster a weight on the basis some aspects. (4) Clusters with a weight greater than a certain threshold are associated to relevant places.

neighboring cells. Other approaches recognize this issue and focus on clusters of cells spanning a given area [1,8]. Our clustering mechanism uses a spatio-temporal distance function to better identify areas that are visited by the user (see Section 4.3).

2. **Weighing.** While most of the approaches just count the number of CDRs or the number of days with CDRs to determine the "importance" of an area, we refine such a classification by combining different measures together. We use weights to take into account the time of CDRs, the number of days, and how evenly CDRs are spread across the week (see Section 4.4).
3. **Thresholding.** While most of the approaches assume the presence of a single home and a single work place for each user, our approach tolerates multiple homes, work and other kind of places, or even none of them. We developed a flexible thresholding scheme to identify relevant clusters (see Section 4.5).

In addition, the majority of the approaches at the state of the art focuses on home and work locations only. They try to identify home locations on the basis of CDRs generated at night, while work locations on the basis of CDRs generated during the day [5,6,1,8,7]. Such approaches use a fixed threshold to identify night and day time intervals. In our approach, we use a more flexible approach and try to extract also other kind of places.

In the following of this paper we first motivate the present work, describe applications scenarios and associated privacy issues (Section 2). Then we present related work in the area and show how our proposal improves over such existing approaches (Section 3). We describe in detail the proposed methodology (Section 4), and we present analysis and results' evaluation (Section 5). Finally we present some concluding remarks.

2. Motivations and privacy issues

The automatic identification of users' places is a relevant research topic in mobility data analysis and enables a number of application and services: (i) extracted places could be the basis to understand mobility patterns in the area. Specifically, from home and work places it is possible to automatically infer recurrent commuting trips. Then, aggregating those trips it is possible to obtain origin–destination matrices and traffic flow over an area [3,1]. This naturally supports urban planning and traffic management. (ii) Knowledge of recurrent places can enable novel services targeting people visiting specific areas [4]. For example, it is possible to personalize outdoor advertising and information billboards on the basis of the places visited by the users in that area. Similarly, an application (like TripAdvisor.com) could rank places on the basis of the number of people routinely frequenting those places. More in general, such a knowledge could enable novel services in smart city scenarios [9]. (iii) The knowledge of places visited by people (i.e., the typical distribution of people in the city) could be also very important in emergency and disaster-response situations to prioritize interventions and manage mobility on the basis of the actual people distribution [10].

Telecom operators are uniquely positioned to extract such information from cellular network data. However, despite this potential for innovation, the analysis of CDRs to extract places frequented by people raises many concerns about individuals' privacy, rightfully constraining their use. In the European Union, the legal framework ruling on this kind of analysis (one of the stricter in preserving individuals' privacy) is described in the *EU Data Protection Directive (Dir. 95/46/EC)* and in the *Article*