Engineering Applications of Artificial Intelligence 54 (2016) 49–61

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Identifying user habits through data mining on call data records

Filippo Maria Bianchi [a,*], Antonello Rizzi [a], Alireza Sadeghian [b], Corrado Moiso [c]

[a] Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy
[b] Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON, Canada M5B 2K3
[c] Future Centre Department, in Telecom Italia, via Reiss Romoli 274, 10148 Torino, Italy

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
19 December 2015
Accepted 15 May 2016
Available online 24 May 2016

Keywords:
Call data records
Clustering
Data mining
Knowledge discovery
Automatic semantic interpretation
Frequent substructures miner
Subspace clustering

ABSTRACT

In this paper we propose a frameworks for identifying patterns and regularities in the pseudo-anonymized Call Data Records (CDR) pertaining a generic subscriber of a mobile operator. We face the challenging task of automatically deriving meaningful information from the available data, by using an unsupervised procedure of cluster analysis and without including in the model any *a priori* knowledge on the applicative context. Clusters mining results are employed for understanding users' habits and to draw their characterizing profiles. We propose two implementations of the data mining procedure; the first is based on a novel system for clusters and knowledge discovery called LD-ABCD, capable of retrieving clusters and, at the same time, to automatically discover for each returned cluster the most appropriate dissimilarity measure (local metric). The second approach instead is based on PROCLUS, the well-know subclustering algorithm. The dataset under analysis contains records characterized only by few features and, consequently, we show how to generate additional fields which describe implicit information hidden in data. Finally, we propose an effective graphical representation of the results of the data-mining procedure, which can be easily understood and employed by analysts for practical applications.

1. Introduction

Thanks to the popularity and wide diffusion of cellular phones, a huge quantity of mobile devices are moving everyday with their human companions, leaving tracks of theirs movements and their everyday habits. Mobile phones are becoming pervasive in both developed and developing countries and they can be a precious source of data and information, with a significant impact on research in behavioral science (Berry, 2011; Lazer et al., 2009).

A Call Data Record (CDR) is a data structure storing relevant information about a given telephonic activity involving an user of a telephonic network. A CDR usually contains spatial and temporal data and it can carry other additional useful information. Population census have been widely used in the past for keeping track of the demography and geographical movements of the population. Nowadays, due to short term and everyday mobility, more flexible methods such as various registers and indirect databases are employed: CDRs represent an optimal candidate in this sense. One of their main advantage is that they offer a statistically accurate representation of the distribution of people in an area and they can be used to track large and heterogeneous groups of people. Since CDRs evolve accordingly to the changes of users behavior, the information they carry "automatically" updates over time. Telecom operators continuously gather a huge quantity of CDRs, from which it is possible to extract additional information with low additional costs and generate valuable datasets. Analyses of CDR data can be successfully employed in many different fields, like monitoring the network, adaptation of supplied services (e.g., customers' billing, network planning), understanding of the economic level of a certain area, performing socioeconomic studies oriented to marketing and to build social networks (Duong et al., 2010). For example, once the relationship between behavior, response, risk or other attributes is established, targeted offers of appropriate products or services can be addressed to specific customers by the telephone companies. Mobile positioning is a valuable source of information for investigating the spatial dynamics of human communities, but the number of published studies on this topic is still poor, mainly because of problems concerning limited access to such data and privacy issues. Localization procedures relying on mobile positioning generally provide less accurate information than the GPS (Global Positioning System), but the latter needs to be turned on to register the position, with a consequent increment of battery consumption. The wide diffusion of mobile equipment, in addition to the widespread installation of radio transmitters in both urban and rural areas, makes such positioning techniques very appealing for many case-based reasoning applications (Olsson et al., 2004). The cellular

* Corresponding author.
E-mail addresses: filippomaria.bianchi@uniroma1.it (F.M. Bianchi), antonello.rizzi@uniroma1.it (A. Rizzi), asadeghi@ryerson.ca (A. Sadeghian), corrado.moiso@telecomitalia.it (C. Moiso).

http://dx.doi.org/10.1016/j.engappai.2016.05.007
0952-1976/

network consists in a set of base stations formed by one tower and several directional antennae. The radio coverage of a single antenna represents a cell, whose size is not fixed in the whole network. Mobile phones can be seen as a wide-area sensor network, whose measurements can be integrated with heterogeneous sources (Bianchi et al., 2015c). GSM (Global System for Mobile communications) is a mobile network based on the communication with an antenna covering a local area; the active connection to a certain antenna represents a spatio-temporal information which can be used for tracking the activity of a user in a GSM covered area. An effective approach for the analysis of CDR is offered by data mining techniques based on pattern recognition and machine learning procedures, which in the last years have been successfully employed in many different fields (Taormina and Chau, 2015; Bianchi et al., 2015b; Zhang and Chau et al., 2009; Wang et al., 2015).

A generic CDR relative to a telephone activity of a subscriber in a mobile network contains the identifiers of the two parties (the one which issued the call and the one which received it), the personal data of the user (name, age, sex, residence address), the coordinates of the cell which served the call, the time when the activity is registered, information on the mobile devices and on the telephonic plan. However, many of these fields are obfuscated or deleted from the publicly available records, in order to protect the privacy of the subscribers.

In fact, digital traces left by mobile phones reveal personal, often sensitive, information about their users. CDR analysis must be ruled according to the privacy of the national regulatory framework. Specifically, CDRs can be collected and processed by a mobile operator to implement the features necessary to deliver the mobile service (e.g., billing, customer care, network operation and planning). Any further usage must be either explicitly authorized by customers (e.g., through consent) or by the privacy authority. In general, the processing of anonymized CDRs (i.e., records in which the personal identification information of the referenced people is removed) is allowed, as these records are no longer personal data. Moreover, under some conditions, there is greater flexibility in processing and exploiting pseudoanonymized CDRs (i.e. records in which the ID of the referenced people is replaced with a code, often obtained through one-way cryptography): in particular, the pseudoanymization procedure must prevent the re-identification of a person from the analysis of the pseudoanonymized records. These conditions set some constraints on the datasets that a mobile operator can analyze or transfer to third parties. Therefore, also in case of pseudoanonymized records, the dataset must be pre-processed in order to reduce probability of re-identification. A common procedure consists in decreasing time resolution or increasing space granularity, so that the data collection never spans long time periods or the spatial information is not detailed. Typical examples are datasets of CDRs with high spatial resolution containing records of users which are monitored for a short period of time, or datasets where user activities tracked for longer time intervals usually come with a lower space resolution. This latter is the case considered for this study and it will be discussed in details in Section 3.

In this paper we propose a data-mining procedure for automatically identifying the recurrent patterns in the telephonic activity of mobile network users, in order to understand and describe their habits. We design an inference system that uses a cluster-based approach to discover regularities among data. Cluster analysis can be framed into unsupervised learning, which is the task of identify hidden structures in unlabeled data. As in our case of study, the ground truth of the expected result is unknown and there is no error or reward signal to evaluate a potential solution. Cluster-based approaches have been successfully applied for the discovery of new concepts in streams of data (Spinosa et al., 2007; Japkowicz, 1999). However, the outcome of the clustering procedure is strongly influenced by the dissimilarity measure adopted and, in general, it depends on a set of configuration parameters. The procedure of tuning such parameters could be difficult and it may requires *a priori* information not always available. As the core inference engine for our application, we use the LD-ABCD algorithm, which has been recently developed by the authors of this paper. LD-ABCD implements a novel cluster-analysis procedure, which has been presented in Bianchi et al. (2015a). In order to validate the effectiveness of the proposed approach, the algorithm has previously been tested on synthetic and benchmarking datasets, where the ground-truth was known. In this work, we integrate our newly-designed system into a larger framework, which has been specifically designed to deal with a novel real-world case of study. Since we do not possess a supervised information on the results, we evaluate the performances of our system through a comparison with a well-established subspace clustering algorithm. The main contributions of this work are summarized in the following:

1. We propose a cluster-based approach for retrieving multiple groups of CDRs, which are similar according to different subsets of features. We do not make assumptions in advance on which characteristics should be taken into account for identifying clusters, or on the total number of clusters. Moreover, each cluster is characterized by its own dissimilarity measure parameters, according to the concept of *local metric learning*. We interpret the well-defined clusters that have been identified as relevant patterns in the activity of a given user. Such patterns are used to generate a *digital fingerprint* representing user's habits, in terms of telephonic activity, geographical movements, time periods when daily communication activities are more frequent, most visited places (home, workplace etc.) and a social profiling. These fingerprints can be employed for different purposes, like profiling and definition of classes of users, depending on the specific application. With respect to other works focused on the analysis of CDR, we propose a new framework based on a complex data mining and knowledge discovery procedure. We show how meaningful patterns can be extracted and used to characterize a user, preserving his privacy and without making any *a priori* assumption on the nature of the data.
2. When data characterized by a high number of distinct features are available, many informative, significant and useful information can be easily derived, more complex analysis can be performed and non-trivial relationships among data can be discovered. However, in our work we process a dataset of pseudo-anonymized CDRs where each entry contains only a limited number of attributes. The problem we face is challenging since it seems that, at a first glance, only naive regularities in the data can be retrieved. In this paper we show how to extract implicit information from the data and we use them for identifying hidden frequent patterns which lead to meaningful results and considerations.
3. We propose an effective method of visualization, which encodes data and information into visual objects. Our main goal is to communicate information clearly and effectively through graphical tools, in order to express and to quantify the results, through visual human interfaces (Fayyad et al., 2002).

The remainder of the paper is organized as follows: in Section 2 we review some relevant works and applications focused on the analysis of CDRs. In Section 3 we present the dataset considered for the analysis, discussing the representation of the data and how implicit information contained in the CDRs can be extracted. In Section 4 we propose a framework that can be used for