Semantic Interpretation of Mobile Phone Records Exploiting Background Knowledge

Zolzaya Dashdorj[1,2,3] and Luciano Serafini[3]

[1] Semantic & Knowledge Innovation Lab, Telecom Italia, Italy
[2] Department of Information Engineering and Computer Science, University of Trento, Italy
[3] Data & Knowledge Management Unit, Fondazione Bruno Kessler, Italy
dashdorj@disi.unitn.it, serafini@fbk.eu

Abstract. The increasing availability of massive mobile phone call data records (CDR) has opened new opportunities for analyzing and understanding real-life social phenomena and human dynamics. In order to better interpret this enormous amount of data it is useful to contextualize them with information about the circumstances under which they has been generated. Nowadays, linked open data initiative provide access to a huge amount of geo-time referenced knowledge about territory and events that happen in the territory. These informations can be used to characterize the aforementioned context. The aim of this Ph.D is to investigate on the intercorrelations between CDR, contexts, and human behaviors. The ultimate goal is to build a stochastic model, that can be used to predict semantic (qualitative) behavioral patterns on the basis of CDR traffic and context and identify and explain anomalous situations on the basis of deviations from standard CDR patterns.

Keywords: telecommunication dataset, stream-data, human activity, human behavior, semantics, ontology, machine learning, knowledge management

1 Relevancy

A huge amount of mobile phone data records (CDR) are generated every day in tracing users phone calls, sms, web surfing, social network interactions etc. These geo- and time-referenced data constitute an important resource of information for investigating on human behaviours. In [7, 4, 10] authors studies individual traits, human mobilities, while [11, 13, 6, 1, 8, 3] predicts communication networks and communication patterns starting from CDR. Currently, most of the analysis generate a quantitative description of human behaviors, presented via visual analytics techniques but they do not provide any insight at the qualitative/semantic level. With the term "qualitative/semantic description of human behaviour" we intend the description of human behaviors in terms of semantically rich concepts (classes and relations of an ontology) which describe, for instance, the actions performed by a person or a group of people, the events they attend, etc. From some pioneering attempts (see e.g.,[14, 1, 7, 4]]) it was clear that inferring semantically rich description of human behaviors from pure CDR is almost impossible; It is necessary to enlarge the analysis including relevant knowledge of the context in which CDR data are generated.

Contextual information includes environmental data (e.g., weather conditions), static description of the territory (e.g., soil destination and points of interests), public and private events (e.g., concerts, sport matches, public spontaneous meeting, strikes, etc) or emergency events (e.g., accidents, traffic jams, etc), transportation schedule, energy or water consumption, etc.

This research aims at discovering the correlations between CDR stream, contexts and human behaviors, and to represent these correlations in a computational stochastic model. Using these model we can realize a set of important tasks such as: characterization of normal or exceptional events, prediction of human activities and events in certain contextual conditions (e.g., during a festival celebration that organized in the center of a city when the weather is sunny or rainy), semantic explanations to the calling or human activity distribution changes. A first attempt to create such a model has been presented in [12][4] within the Orange "Data for Development" challenge [2]. In this paper, after presenting some related work in Section 2 we describe the main thesis objective and the methodology (sections 3–6). Finally, we summarize the preliminary results and the evaluation plan and future research works (Section 8,9).

2 Related Work

The analysis of CDR by new methodologies proposed by the researchers has made great progress in the areas such as emergency response, city and transport planning, tourism and events analysis, population statistics, health improvement, economic indicators, so on [9, 2, 15, 5].

A social response to the events, in particular, behavior changes have been studied by J.P.Bagrow et al [7]. The authors explored a social response to external perturbations such emergency (bombing, plane crash, earthquake, blackout) and non emergency (festival, concert) events in order to identify real-time changes in communication and mobility patterns. The result show that under extreme conditions the level of communications radically increased right after the emergency events occur and it has long term impacts.

In [4] Calabrese et al analyzed the mobility traces of user groups with the objective of discovering standard mobility patterns associated to special events. In particular, this work analyses a set of anonymized traces of the users in Boston metropolitan area during a number of selected events that happened in the city. A result of such an analisys is that users who live close to an event are preferentially interested in that event. Similarly, Furletti et al [1] analyzes human motion associated to specific human profiles; commuter, resident, in-transit and tourist. Users are classified by a neural network, called self organizing map in one of these profiles, and the result is that the percentage of resident was compatible with the customer statistics provided by the Telecom operator. The short-ranged temporal profiles like commuter and in-transit are significantly vary

[4] http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

and distinguishable than the larger extent profiles like resident. This analysis has been done in the city of Pisa and from the users temporal profile they identified a peak that was caused by the earthquake emergency.

Regarding the idea of collecting contextual information in the mobile phone data analysis, we propose, the similar work has been done by Phithakkitnukoon et al [14]. In this work, geographical information (Point of Interest) are collected using pYsearch (Python APIs for Y! search services) from a map. They annotated the POIs with four type of activities; eating, recreational, shopping and entertainment. The authors analyzed human activity patterns (i.e., sequence of area profiles visited by users) correlating to geographical profiles. Bayesian method is used to classify the areas into crisp distribution map of activities, which enables the activity pattern extraction of the users. The results shows that the users who share the same work profile follow the similar daily activity patterns. But not only these area profiles can explain the mobility of these users and the enlargement of activity or event taxonomy in those areas can enable the classicaions of these activity patterns.

3 Problem Statement

In this research, we are interested in analyzing the CDR in order to discover high level human behavioral patterns that can be described in qualitative/semantic terms. In other words, we generate a *semantic description of human behaviors in certain situations* when the mobile network events (phone call, sms, internet connections etc) are occurred. A semantic description of human behaviors is a representation of the behaviors of a single person, the behaviors of a group of people or of the events that happen in the human society, in terms of concepts and relations of an ontology describing human behaviors and events. An example of semantic description of human behaviors is the fact that "a person is performing some specific actions" (e.g., working, shopping, hiking), or the fact that "certain events are happening in a certain area" (e.g., a car accident, a train suddenly stops in the middle of nowhere, etc.)

To infer these types of information about human behaviors from the CDR, we need to complement these data with contextual information, which describe the context where the mobile network events are occurred. The context is a pair $\langle l, t\rangle$ where l is a location (= geographical area) and t is a time interval. For every context $\langle l, t\rangle$ from one or more knowledge repository we can extract $K_{l,t}$ which is a knowledge base describing this context. $K_{l,t} \equiv O_l \cup E_{l,t}$ i.e., it is the union of the objects which are nearby the location l (point of interests), and the events which take place nearby l at time around t. Every element of O_l is a pair $\langle poi, w\rangle$ where $w \in [0, 1]$ is a weight that expresses the closeness of poi to l. Every element of $E_{l,t}$ is a pair $\langle e, w\rangle$ with $w \in [0, 1]$, expresses how close to l and t is the event. Examples of POIs are buildings, roads, natural points, shops, etc., examples of events are weather phenomena (rain, snow, etc.) or social events, like concerts, strikes, traffic jams, etc. For every context, from the knowledge repository about the context, we can derive $A_{l,t}$ which are the most probable human activities in the context $\langle l, t\rangle$. Every element of $A_{l,t}$ is a pair $\langle a, w\rangle$ where

$w \in [0, 1]$ is a weight that expresses the likelihood of a parson performing the activity a in the context $\langle l, t\rangle$. Examples of the activities are working, studying, shopping, attending in a concert, travelling by car, etc

Making use of the knowledge repository about contexts, we enrich the CDR with human activities and events. Contextually enriched CDR can be exploited to analyse and evaluate call patterns associated to human activities and events. A call pattern is a quantitative model that describes the "normal" behaviour of a communication in a certain class of contexts. An example of a call pattern is the function that associates the number of calls done by people before, during, and after a particular event (e.g., a football match, or a concert). Another example of call pattern (also called interaction pattern) describe the number of calls between a pair of locations before, during one particular time of the working day. A third example of call pattern (also called mobility pattern) describes the number of user displacement from one location to another. Call patterns describes the normal behavior patterns. Comparing call pattern with the actual calls allows us to identify divergent and exceptional behaviors and context allows us to characterize such behaviors (e.g., offering explanations for these behaviors) and the prediction of similar patterns.

4 Research Questions

A more concrete formulation of the research questions are presented below. The questions are posed to discover the correlations between CDR stream contexts and human behaviors and to represent these correlations in a computational model, that can be queried to obtain the following informations:

RQ1 *What are the correlations between contexts and human behaviors? What are the most probable action that a person is doing in a given context?*

RQ2 *What are the correlations between contexts and CDR? What is the normal call frequency pattern in a specific context (where a context can be, a type of area, an event, a time of the day etc..)?*

RQ3 *What are the correlations between call frequency and human behaviors? What is the call frequency pattern of people while performing certain actions?*

To achieve the answers for these core questions, we need to engage the preliminary challenges in linking the data coming from different datasources and real-time knowledge reasoning and pattern recognition in streaming data from the view of computational and conceptual perspectives.

5 Hypotheses

The main hypothesis to this Ph.D research work is to characterize a human behavior that represented in the form of call frequency pattern as it is connected to the semantics of human activities or events in a certain context. The multi-classification of these human behaviors recognizes the situation changes that engaged in the context (e.g., when the weather is sunny or rainy in a certain location and time). This improves the prediction task of human behaviors.

6 Methodology

We organize the work for the semantic interpretation of human behavior in mobility based on the merge of mobile network data stream and the geo and time referred available background knowledge in the following phases:

1. The starting point is a characterization of the territory with the human activities and events that can be performed in any location of an area and at any time of a day. This phase is intended to answer the RQ1 described in Section 4.
2. On the basis of the correlation obtained in phase one, we annotate mobile phone network events with the human activities or events and then extract the semantic behaviors determining the typical calling patterns that associated to various typical activities or events. This answers the RQ2,3.
3. The result of the second phase can be used to explain the possible or anomalous actions and situations in real-time, when the call activity sensibly deviates from the standard call activity associated to a known context.

6.1 Geographical Area characterization

We characterize the territory with contextual information in which mobile network events are occurred, in order to model the relation between human activities and contexts. All the possible contextual information can be retrieved through the employment of online and offline techniques in information retrieval and text mining, probability inference and those information can be scored in order to determine the importance. The example of contextual information include mobile cell covereage map, POIs distribution, social event distribution and domain statistical data about demography, ethnography, energy or water consumption, so on. The allocation of contextual information to each context enables analysis and identification of the possible human activities or events associated with a likelihood which expresses a probability of the activity that could be performed by the users. For example, an area which contains mostly about highway and if there is an accident on the highway, the probable event is a traffic jam while people are performing an action, "travelling by car". For modelling the relation between contexts and human activities, we propose two steps; (1) an ontological model that describes the concepts and the relations between the ontologies of human activities and knowledge about contexts, under the expert knowledge derived from surveys, crowdsourcing and domain experts (2) a stochastic behavior prediction model that predicts the possible top-k activities or events associated with a likelihood that could be performed in each context. The approach quantifies the correlation between contexts and human activities through uncertain probabilistic modeling techniques such as Probabilistic Boolean Networks, Markov Logic Networks, Bayesian Networks. The approach can enable a further consideration of the OWL language extensions with the probability of the activities or events in order to do reasoning with OWL for the prediction.

6.2 Extraction of Semantic Behavioral Patterns

On the basis of the previous association, context → human activities, in this step, we propose a semantic behavioral analysis model that annotates the CDR to the most probable actions/events that happening when the phone calls are localized in the territory. By adapting the state of the art techniques of behavioral call frequency pattern extraction in the area of mobile phone data analysis, we extract a standard type of call frequency patterns about human mobility, communication and interaction patterns that annotated with the human activities/events. The extracted semantic behavioral patterns are classified into certain types based on the similarity metrics of the call frequency patterns which characterized with the contextual knowledge, making use of classification techniques such as Logistic regression, Naive bayes, Perceptron, SVM, and novel classifier fusion methods so on. The classifcation results are stored into a behavioral decision tree repository in each area. This explains the correlation between CDR and contexts as well as human activities or events. Example of semantic behavioral pattern is attending in a concert can be vary depends on the weather condition, geographical location and the events that could be occurred at the same time, so on.

6.3 Forecast of the CDR stream in Real-time

Exploiting those models in real-time, we will forecast the CDR stream in the given territory to explain the possible or anomalous actions and situation changes in real-time. We adapt algoritms which operate in online and incremental fusion such as streaming linked data framework, C-SPARQL that will propose online techniques for annotating the semantic labels of contextual information that described in the form of Linked Open Data with the call frequency patterns in CDR stream. The CDR stream can be transformed into RDF stream and the reasoning of the stream can be done. The observed call frequency patterns are analysed and reasoned comparing to the classification of semantic behavioral patterns that stored in the behavioral desicion tree repository in each area. This enables identification and prediction of standard or anomalous type of behaviors in real-time offering semantic explanations to the CDR stream. The new discovered behaviors can be learned to the behavioral desicion tree repositories.

7 Reflections

This Ph.D research work is intended to develop a novel model that deepens the analysis of the CDR through machine learning approaches and logical semantics considering a wide range of contextual features in each context where mobile network events occur. This could provide an extensive overview to the CDR and that could be interpreted in qualitative terms.

8 Evaluation Plan

An evaluation will be divided according to each phase of the methodology that described in Section 6:

Phase1 At the first phase, we evaluate the model that quantifies and qualifies the correlation between human acivities and contexts based on user data (ground truth) we have collected through a web or mobile phone application, about daily activities that performed in various areas of the territory and in different times of a day. We choose several cities as use case in order to do comparative analysis. The preliminary evaluation has been done. In this evaluation, the model charaterizes every context of the mobile network events in the Trento city, Italy with the possible human activities that extracted from a geo-referenced datasource, OpenstreetMap. By collecting user-feedback, we obtained 70.89% of overal accuracy, and 61.95% of overal accuracy among the top-5 activities.

Phase2 At this phase, the evaluation is concentrated in the correlation between CDR and contexts as well as human activities. We use a sample CDR dataset (training, test) that covers particular events and festivals, concerts, etc that selected for the evaluation. In training dataset, the semantic behavioral patterns are extracted and classified into certain types.. The test dataset is used to measure the accuracy of the performance of the classfication model for predicting semantic behavioral patterns considering the call frequency patterns associated with similar events, festivals and concerts, etc. We compare the results in different cities.

Phase3 We concentrate in the real-time analysis of the CDR stream to evaluate the performance of identification and prediction of possible or anomalous actions and situations in real-time. We use the training dataset which used in Phase2 in order to use the semantic behavioral patterns for identification and prediction of actions in the CDR stream. We evaluate the resuls with a help of domain experts.

9 Preliminary Results and Conclusion

In this paper, we presented the Ph.D work aimed at understanding the correlations between CDR, human behaviors, and contexts in a computational model. To understand these correlations from quantitative data (CDR), we complement contextual information in order to describe the context where a phone call is done. Our methodology which addresses these problems is divided into three core phases 1) geographical area characterization 2) extraction of semantic behavioral patterns 3) forecast of the CDR stream in real-time. At the first phase of the evaluation, we obtained 70.89% of overal accuracy, and 61.95% of overal accuracy among the top-5 activities.

Next step of this Ph.D work is to enrich the geographical area characterization making use of various type of geo/time-referenced contextual information available on the web sites such as environmental data about weather condition, and public and private events about festivals and concerts and emergency events about accident or strike and some other domain statistical data about energy consumption, so on. We organize a wide range of evaluation for this model, involving as many as participants who can share their daily activities. On the basis of this model, we will work on the following phases; extraction of semantic behavioral patterns and forecast of the CDR stream in real-time that will allow us to determine semantically rich heterogeneous (normal or anomalous) human behaviors in real-time.

References

1. B.Furletti, L.Gabrielli, C.Renso, and S.Rinzivillo. Identifying users profiles from mobile calls habits. In *the Proc. of the ACM SIGKDD Int.Workshop on Urban Computing*, UrbComp '12, pages 17–24. ACM, 2012.
2. Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.
3. Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.
4. F.Calabrese, F.C.Pereira, G.Di Lorenzo, L.Liu, and C.Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th Intl.Conf. on Pervasive Computing*, Pervasive'10, pages 22–37, 2010.
5. Ferrari.L, Berlingerio.M, Calabrese.F, and Curtis-Davidson.B. Measuring public-transport accessibility using pervasive mobility data. *IEEE Pervasive Computing*, 12(1):26–33, 2013.
6. J.Candia, M.C.González, P.Wang, T.Schoenharl, G.Madey, and A.Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, June 2008.
7. J.P.Bagrow, D.Wang, and A.Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.
8. R. Lambiotte, V. Blondel, C. Dekerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, September 2008.
9. K Laurila.J, Gatica-Perez.D, Aad.I, Blom.J, Bornet.O, T. Do, Dousse.O, Eberle.J, and Miettinen.M. The mobile data challenge: Big data for mobile computing research. Newcastle, UK, 2012.
10. M.C.Gonzalez, C.A.Hidalgo, and A.Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.
11. J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.
12. P.Paraskevopoulos, T.Dinh, Z.Dashdorj, T.Palpanas, and L.Serafini. Identification and characterization of human behavior patterns from mobile phone data, 2013.
13. Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.
14. S.Phithakkitnukoon, T.Horanont, G.Di Lorenzo, R.Shibasaki, and C.Ratti. Activity-aware map: identifying human daily activity pattern using mobile phone data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.
15. A. Wesolowski, N. Eagle, A.J. Tatem, D.L. Smith, A.M. Noor, R.W. Snow, and C.O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–70, 2012.