Semantic Enrichment of Mobile Phone Data Records Using Linked Open Data

Zolzaya Dashdorj, Luciano Serafini

SKILL, Telecom Italia and DKM, Fondazione Bruno Kessler
University of Trento, ICT International Doctoral School, Trento, Italy
{dashdorj@disi.unitn.it,serafini@fbk.eu}

1 Introduction

Users of mobile (smart) phones, generate an enormous amount of data every day. Most of them are not accessible due to privacy reason, but anonimized metadata, such as for instance, the location, the time and the duration of the interaction with the smartphone are nowadays available for analysis. We address these data as "Call Data Records" (CDR). CDR metadata constitute an important source of information for investigating on general human behavior, such as mobility [5, 3, 6], and communication patterns [9, 2, 4, 1, 8]. Currently, most of the analyses provide a quantitative description of human behavior, which is presented via visual analytics techniques. The outcome of these analysis are usually quantitative models estimating for instance, the number of people present in a certain area at a certain time, the number of people who moves from point a to point b within a certain period, and so on. Less interest has been dedicated to the creation of model that describe human activity at qualitative/semantic level, i.e., in terms of semantically rich concepts in order to estimate/predict for instance, the actions performed by a group of people in a certain situation or the type of event is happening around in a certain location, on the basis of the CDR.

The analyses presented in [11, 10, 1, 5, 3] need to be extended making use of a relevant knowledge of the context of the user. User contextual information includes all those information describing the objects present and the events happening in the place and at the time that the user is interacting with his/her smartphone. Contextual information includes information about the territory (e.g., points of interests - POIs), weather conditions, public and private events (e.g., concerts, sport events, public spontaneous meeting etc) and emergency events (accidents, strikes, etc.), transportation schedule, energy or water consumption, etc.

In this paper, we propose a first step of investigation developing an ontological and statistical model (HRBModel) capable to predict the possible human activities in a certain place and on a certain time on the basis of the contextual information describing the POI's of the place and information about the time of the day at which certain actions are usually performed. POI's are taken from Openstreetmap[1]. The model enables early identification of standard type of hetergenous human activities in various geographical area profiles and at different times in which CDR occur.

2 Experimental Setup

We are interested in predicting the most probable human activity of a user when he/she is in a specific geographical area at a given time. We develop our experiment consider-

[1] http://www.openstreetmap.org/

ing the area of Trento - Italy, but the process and the software is a general enough to be applicable to any geographical area.

Details about the experiment are described in the following.

We divide the geographical area of the city into sub-divisions in a way that POIs are uniformly distributed over the subareas (see Figure 1(a)). In each sub area, POIs are extracted using geographical tools like OSM2PGSQL. We do not consider the POIs which don't derive human activities, such as benches, towers, emergency phones, etc. In total for this city, we extracted 333,809 POIs from the OSM map which cleaned to 159,314.

To annotate the POIs with the human activities, we propose a Human Behavioral Ontology (HBOnto) that with the help of the OSMOnto ontology[7], associates POIs with all the human activities that can be performed or hosted there or nearby. The human activities are hierarchically organized from specific activities to 10 high level categories, that around 220 human activities. For this association, we take day and time into account as all the activitiies are connected to a day-time range of validity. A stochastic behavior model (SBM) we propose that estimates the probability of human activities given the location and time of an event, on the basis of the ontological model as follow:

$$\mathrm{P}(a|t,l) = P(a|t) * P(a|l) \text{ (t, l are independent)} \tag{1}$$

As an initial step, we manually created a fuzzy model for $P(a|t)$ that performs reasoning on the importance of activities given a time that is estimated as follow:

$$\mathrm{P}(a|t) = \frac{FuzzyActivity(a,t)}{FuzzyTime(t)} \tag{2}$$

$P(a|l)$ is estimated based the importance of the activities given a location using TF/IDF for the POIs importance that derives a weight to the activities as follow:

$$\mathrm{tf-idf}(f,l) = \frac{N(f,l)}{\underset{w}{\mathrm{argmax}}\{N(w,l) : w \in l\}} * \log \frac{|L|}{|\{l \in L : f \in l\}|} \tag{3}$$

To avoid the spatial gap, $P(a|l)$ can be extended if we consider the nearby activities in a given radius r of a circular area around the location.

$$P(a|l,r) = \frac{\underset{a}{\mathrm{argmax}}\{W(a,l_i) * \lambda_i\}}{\sum\limits_{a \in r} \underset{a}{\mathrm{argmax}}\{W(a,l_i) * \lambda_i\}} : r \bigcap_{i=1}^{n} l_i \tag{4}$$

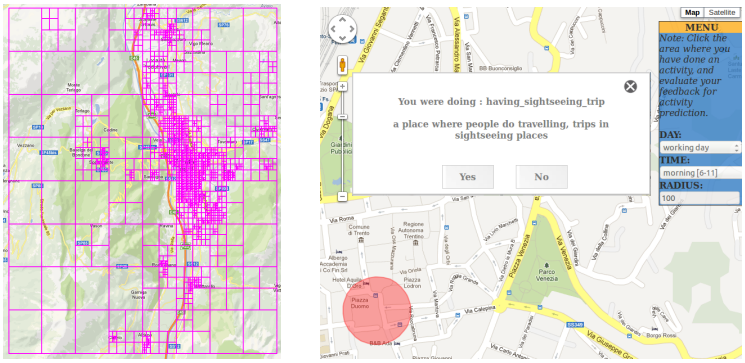
We collected the user-data[2] through the experiment application described above (see Figure 1(b)) for one week with 32 partipants involved (see Table 1). It emerged that most of the user feedback comes from the areas of (Trento-Povo and Trento-Downtown).

Every user's feedback is collected in a record containing: the latitude/longitude of the location selected on the map, the radius of the area, the selected activity (among top-5 or one freely chosen), the semantic day and time (e.g., weekday, saturday.., early morning, mid morning.. etc), and the current time of the system.

[2] http://dkmlab.fbk.eu:8080/BHRModelTest/data/semantic_data_BHRModel2013.csv

Number of feedback	Participants	Duration	Feedback clusters	Feedback in each cluster
481	32	1 week	5	Trento.Nord - 21, Downtown - 180, Povo - 151, Santa Chiara - 60 Trento.Sud - 67
Feedback on Weekday		morning 158, mid day 29, afternoon 59, evening 61, night 28		
Feedback on Saturday		morning 19, mid day 5, afternoon 33, evening 19, night 8		
Feedback on Sunday		morning 22, mid day 5, afternoon 25, evening 6, night 4		

Table 1: Data collection by user feedback over different locations and times



(a) Geographical area divisions by the density of POIs

(b) Visualization of demonstration app, at Piazza Duomo on weekday mornings

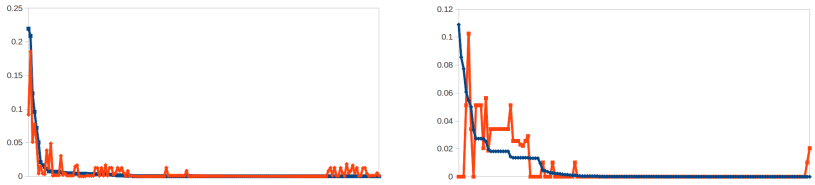
Fig. 1: Application UI for High level Behavioral Representation Model

3 Preliminary results

Given the collected data, we measured the accuracy of the HRBModel considering the correctly predicted activities w/regardless of the ranking position. We also analyzed the divergence of the probability activities comparing to the probability from the feedback in the areas with the highest number of feedback: Trento-Downtown and Trento-Povo. The example of the results, Figure 2(a) and 2(b) describe the divergence of the probability activities that occur in those areas on weekday mornings and afternoons, in which we propogated the probability activity to the child activities. In these figures, the activity indexing (x-axis) is different in each area, y-axis is the probability associated to the activities. The figures show that the probability activities from the user feedback can still follow the trend of the probability from our model.

4 Conclusion and Future works

Within the 481 users' feedback collected, 341 activities have been correctly predicted, corresponding to an overall accuracy of 70.89%. The overall accuracy of a correct activity prediction (among the top-5) corresponds to 61.95%. We have done the evaluation considering the high level (parent) activities, the overall accuracy has been increased to 80.23%. We showed the divergence between the probability activities in our model compared to the probability from the feedback by various locations and times that can be further studied in order to understand the correlation between human activities and contexts. We will extend the evaluation of the model making use of mobile phone survey,



(a) Trento-Downtown area on weekday mornings

(b) Trento-Povo area on weekday afternoons

Fig. 2: Divergence of the activities probability from our model (blue) and from user feedback (red).

crowd sourcing, and social networks (e.g., twitter, foursquare). The proposed model will be a baseline to characterize geographical areas by activity of interests in the areas where CDRs are occurred. This allows identification and prediction of human behaviors by various area profiles (e.g., business, shopping, or leisure areas etc) in certain contextual conditions and detection of anomalous behavioral events.

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. C.Ratti, S.Sobolevsky, F.Calabrese, C.Andris, J.Reades, M.Martino, R.Claxton, and S.H.Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.
3. 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.
4. 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.
5. J.P.Bagrow, D.Wang, and A.Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.
6. M.C.Gonzalez, C.A.Hidalgo, and A.Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.
7. M.Codescu, G.Horsinka, O.Kutz, T.Mossakowski, and R.Rau. Osmonto - an ontology of openstreetmap tags. In *State of the map Europe (SOTM-EU)*, 2011.
8. N.Eagle and (Sandy) A.Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.
9. J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.
10. 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.
11. Z.Dashdorj and L.Serafini. Semantic interpretation of mobile phone records exploiting background knowledge. In *Intl.Conf. Semantic Web Conference (ISWC 2013), Doctoral Consortium*, 2013.