ABSTRACT
The pervasiveness of mobile phones creates an unprecedented opportunity for analyzing human dynamics with the help of the data they generate. This enables a novel human-driven approach for service creation in a variety of domains (e.g., healthcare, transportation, etc.) Telecom operators own and manage billions of mobile network events (Call Detailed Records - CDRs) per day: interpreting such a big stream of data needs a deep understanding of the events’ context through the available background knowledge. We introduce an ontological and stochastic model (HRBModel) to interpret mobile human behavior using merged mobile network data and the geo-referenced background knowledge (e.g., OpenStreetMap, etc.). The model characterizes locations with human activities that can happen (with a given likelihood) there. This allows us to predictively compile sets of tasks that people are likely to engage in under certain contextual conditions or to characterize exceptional events detected from anomalies in the CDR. An experimental evaluation of the approach is presented.

Categories and Subject Descriptors

General Terms
Algorithms, Human Factors

Keywords
mobile phone data records, human activity recognition, context aware computing, qualitative methods, human behavior, semantics, ontology, machine learning, knowledge management, Linked Open Data

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

MUM 13, December 02 - 05 2013, Luleå, Sweden.
Copyright 2013 ACM 978-1-4503-2648-3/13/12 ...$15.00.
http://dx.doi.org/10.1145/2541831.2541857

Mobile phones are the most pervasive personal device ever they became part of a daily life. All mobile network events, such as calls, SMS and data connections are traced in logs, which offer a great potential to understand where people are and what people are doing. Researchers in the areas of behavioral and social sciences are interested in examining Call Data Records (CDR) to characterize and understand real-life phenomena, such as individual traits, human mobility [1, 3, 5, 17], communication and interaction pattern[3, 12, 16]. All these approaches make use of models for human behavior to provide quantitative evidence (graphs, tendencies, etc.). In this paper, we are interested in a qualitative description of human behaviour as opposed to the current trend of considering numerical aspects derived from the collected data. A qualitative description of human behaviour is given in terms of semantically rich concepts that refer to an ontology of human actions and events and to environmental descriptions. However, without knowing the context where the CDR are generated human behaviors are not easily recognized. The context is a pair of location and time, where location is a geographical area and time is a time interval. For every context, we can extract the relevant background knowledge describing it. In some previous work (see [10, 1, 3]) a classification of user profiles and event types is inferred from the CDR analysis. However, we argue that a much more effective interpretation of the CDR stream is obtainable by using a background knowledge that helps in describing the contexts of the network event thus resulting in an enhanced behavioral pattern understanding [9, 8].

As an improvement to previous works( see [15, 10, 1, 3]) sharing the same purposes, the interpretation of the mobile network events needs to be extended considering a relevant knowledge of the user context. User contextual information describes the objects near the user location and the events that happen there at the time when the user interacts with his/her mobile. Contextual information is available on the web in the form of geo-tagged resources, geotime referenced event information or weather conditions data. Some examples are objects (e.g., buildings, streets, etc.) close to a user, weather conditions (e.g., raining, sunny, etc.) during mobile network events, social events in which a user is participating (e.g., a concert, a meeting, a strike), the type of a day, (e.g., a weekday, a holiday) and so on that potentially can influence the behaviors of the users. The contextual information revealed itself as a great source of improving the qualitative analysis and classification of human behaviors. This research is aimed at studying the correlations between the CDR, the contexts, and the human behaviors.

In this paper, we proposed an ontological and stochastic model (called HRBModel) as a first step to study the correlations between human activities and the contexts where phone calls occur. The
model predicts the possible human activities that can take place in a certain area and at a certain time on the basis of the available background knowledge. Contextual information is extracted from OpenStreetMap (OSM). The model enables early identification of standard types of human activities in various geographical area profiles and times in which CDRs occur. A first attempt to create such a model has been presented in [14] within the Orange "Data for Development" challenge [2]. This paper extends that approach.

The model will allow us to associate the relevant contextual information (e.g., various typical human activities or events) with mobile network events by analyzing the normal calling frequency patterns from the CDR that occur in given contexts. The model is tested through a preliminary experimental evaluation. Further, we can improve the implementation of a set of predictive tasks, such as the prediction of human behaviors under certain contextual conditions (e.g., when an accident occurs on a highway before working hours, etc.) or the characterization of exceptional events detected from anomalies in the mobile network data. In the paper we also describe a preliminary experimental evaluation that measures the performance of the model in predicting the activity of a user.

The paper is structured as follows: Section 2 briefly introduces the related work. Section 4 presents a novel method for the semantic interpretation of human behavior in mobile networks. The design of the experimental evaluation and preliminary results are illustrated in Section 5. Finally, we summarize the discussion in Section 6.

2. RELATED WORK

Furletti et al. [10] extract user profiles from CDR. The authors analyze human movement behaviors, which correspond to specific human profiles such as commuter, resident, in-transit and tourist, that are inferred by profile assumptions. A classification technique, based on a neural network, named a Self Organizing Map is used to classify users by similar call profiles constrained by their temporal distributions. Results show that the percentage of resident was compatible with the statistics provided by the Telecom operator, and short-ranged temporal profiles, like commuter and in-transit are varied significantly than the profiles with a larger extent, like resident. The authors tested their approach on a case study in the city of Pisa with a volume of data consisting of around 7.8 million call records in a month. Analyzing the temporal distribution of calls, the authors recognized a very high peak of calls, all of them concentrated on a short period. They found out that this peak took place immediately of the reporting of earthquake news. Starting from this observation, the authors highlighted the necessity to align temporal call distributions with high level observations concerning events and other contextual information coming from different datasources in order to have a more fine-grained interpretation of the phenomena.

In [3], the authors analyze the mobility traces of groups of users with the objective of extracting standard mobility patterns for people during special events. In particular, this work presents an analysis of anonymized traces from the Boston metropolitan area during a number of selected events that happened in the city. They finally proved that people who live close to an event are preferentially interested in events organized in the proximity of their residence.

Phithakkitnukoon et al [15] analyze the correlation of geographic areas and human activity patterns (i.e., sequences of daily activities). Python APIs for Y! search services, and pYsearch are used to extract Points of Interest (POI) from maps. The POIs are annotated with activities like eating, recreation, shopping and entertainment and then Bayes theorem is used to classify the areas into a crisp distribution map of activities. Based on the user trajectories, the authors identify the work location as the place of frequent stops during the day that derives the user mobility choices by the daily activity patterns. Stop extractions are done the same way as in [3]. The study shows that the people sharing the same work profile have strongly similar daily activity patterns. But this similarity is reduced when the distance of the work location increases. Due to the limitation of heterogeneity of activity, the results can not explain some certain behaviors like night activities in shopping areas.

Social response to exceptional events are studied in [1]. The authors explored social response to external perturbations like bombings, plane crashes, earthquakes, blackouts, concerts, and festivals to identify real-time changes in communication and mobility patterns. Results show that user behaviors under extreme conditions radically changes right after emergency events occur, and their impact is long term.

Calabrese et al. in [3] characterize the relationship between events and their attendees. With an approach based on [4], they determined user trajectories with a specific focus on the places where people stop on for how long they stop. The stop duration of the user trajectory is estimated based on the consecutive calls. Given an event, for each sub-area of the grid, the number of people who are attending that event and whose home location falls inside that sub-area describes the attendance of that in geo-space. Most of the people attending one type of event are most probably not attending other types of event, and people who live close to an event are preferentially attracted by it. As a consequence, the approach can partially predict the initial locations of people who are coming to the future events at that location. This approach could be useful for determining anomalies and additive travel demands for event capacity planning considering the type of events. Moreover, knowing the interests of people in events, the type of an event they are going to be predicted from their mobility patterns. Determining the actual number of attendees as well as the validation of the models are still open problems due to the noisiness of ground truth data. So, it uses other approaches issues like refining mobility patterns belonging to the events, which occurred in the similar region close in time, distinguishing home locations of people who live in same place where events take places.

Regarding the current way of mobile phone data analysis, the need of semantic information about physical, geographical, and social environments, the weather and all the relevant knowledge from various datasources is emphasized in order to perform better analysis[1, 15, 3]. In general terms the contextual information for a mobile network event is all the set of information that contributes to specifying the context (time + area) where the mobile network event has occurred. All the information available on the Internet, which is geo-time referenced can be considered useful for providing a semantic interpretation of the mobile network event for classifying human behaviors. On the practical side, however, we have to concentrate on those sets of resources that are easily accessible and contain enough information for our purpose.

1http://www.openstreetmap.org
3http://pysearch.sourceforge.net/
3. INPUT DATA
In this section, we describe the format of the data on which we have our analysis. There are two types of data; CDR and sources of contextual information; we describe them in the following two subsections.

3.1 Mobile phone data records
The type of a CDR dataset provided by Telecom Italia which is completely anonymized and composed by records with the following structure:

- **source cell and target cell**: the identifiers of the cell from which the call is issued and the receiver’s cell, respectively. They are composed of a Location Area Code and additional parameters such as the operator, etc.
- **date and time**: The date and the time of the day in which the call is issued
- **duration**: The duration of the call
- **source and target cell identifiers**: The identifiers of the cell from which the call is issued and the receiver’s cell, respectively. They are composed of a Location Area Code and additional parameters such as the operator, etc.

Using the cell identifier we can locate the cell on the territory. Indeed each call is associated to a coverage area having a detailed geo-spatial description. A mobile phone cell coverage area is a set of geographic regions associated with the antennas (cells) of the mobile network. Each mobile network event in a CDR is associated with the cell where the user was located. The size of the coverage areas is very assorted depends on multiple factors. However, in remote areas, the coverage can reach up to 35 km². This definitely affects the quality of the analyses in low-density areas. It is observed that, in the presence of regular territory (i.e., flat with no mountains or other natural irregularities), the shape of areas can be approximated by convex polygons. Otherwise, the area can be very irregular and possibly disconnected. The example of the two extreme cases is shown in Figure 1. Irregular coverage areas make the task of estimating the calling location particularly difficult.

![Figure 1: Example of the two extreme cases of cell coverage area](image)

3.2 Geo-referenced objects
One key element in the description of the context of a region is the set of objects that populating it (e.g., a restaurant, an ATM, a bus stop, etc.). These objects provide good indicators for identifying the activities people are more likely to perform in that area.

We get these objects populated in a region using the OSM. OSM is an open and free map of the world with rich geographical objects under an open content license. The database of OSM is frequently updated and is rapidly growing. It has a free tagging system based on a well documented taxonomy, which classifies the objects in categories such as roads, buildings, etc. OSM adopts a topological data structure based on the three core elements:

- **Nodes**: points with a geographic coordinate (i.e., a pair of latitude and longitude). A node mostly describes a POI. For example, shops, restaurants, etc.
- **Ways**: ordered lists of nodes in a poly-line or a polygon. These elements are used for representing linear features or areas, such as streets and forests.
- **Relations**: groups of nodes and ways representing the relations among existing nodes and ways, such as roads with several exiting ways.

The classification of OSM objects is specified via tags\(^4\) (i.e., a textual description captured as a pair of a key and a value). The tags for nodes, ways, and relations are represented as a set of pairs of <key, value>. For instance, OSM tags for a node corresponding to a supermarket are the following:

```xml
<node id="618033185" version="3" uid="330007" user="pikappa79"
timestamp="2011-05-26T16:02:14Z"
changeset="8254868"
lat="46.0946011" lon="11.1162507">
  <tag k="addr:postcode" v="38121"/>
  <tag k="shop" v="supermarket"/>
  <tag k="addr:country" v="IT"/>
  <tag k="name" v="Supermercato PAM"/>
  <tag k="addr:housename" v="Bren Center"/>
  <tag k="addr:street" v="Via Giovanni Battista Trener"/>
  <tag k="addr:city" v="Trento"/>
</node>
```

This is much more effective in future if intersected with time references: a call event in a CDR can be more likely associated with a type of action the calling person is performing. For instance, a person is close to a restaurant and the time is around 1pm or 8pm, then (s)he is probably looking for some food while it may be associated with other activities in other time periods. Our approach is not specific to OSM, and is generalizable by considering resources for instance, Foursquare, Google Map, Yahoo Map, or regional cadastral map databases.

4. METHODOLOGY
We created a High-level Representation Behavioral Model (HRB-Model), making use of: (i) Human Behavioral Ontology (HBOnto) - an ontological model containing an organized set of POIs where each POI is associated with all the human activities that can be performed or hosted there or nearby. Every activity is associated with a day-time range of validity. (ii) a stochastic behavior model (SBM) for associating a measure of likelihood to the activities based on the cell coverage area where mobile network events occurred and the day time of the events.

The HRBModel can be used to predict the most probable activity in a given region at a given time by applying the following three steps: (a) given the region, we obtain its POIs from OSM; (b) from the POIs we can obtain the possible actions in the region and rank them according to their importance and frequency, as well as the temporal information. In the following we provide a detailed description of the elements of the HRBModel.

\(^4\)http://wiki.openstreetmap.org/wiki/Map_Features
4.1 Human Behavioral Ontology (HBOnto)

The Human Behavioral Ontology (HBOnto) is used for modeling the possible human activities in terms of POIs and days, times. It is composed of three models (Figure 2(a)), one describing the POIs and their relations (OSMonto), another containing the temporal references (TimeOnto), and the third (ActOnto) describing the human activities belonging to POIs. HBOnto links POIs types and the time of the day to human activities. Every POI type is related with a set of activities \( A(\text{poi}_i) \subseteq \{a_1, a_2, \ldots, a_n\} \). Those activities will be associated with probability \( p(a_1), p(a_2), \ldots, p(a_n) \). There is a validity boolean function \( \text{VAL}(\text{poi}_1, a_1, t_1) \) that indicates whether an activity is valid at \( t_1 \) or not.

We offer the details on the three ontology below.

### 4.1.1 ActOnto

Having the POIs in the OSMonto, we identified a set of human activities that can be performed or hosted within every POI or nearby. It is the case, for instance, of EATING in a RESTAURANT or TRAVELING on a RAILWAY. In the ActOnto model, we hierarchically organized 222 human activities into 11 categories. We use description logic to define the axiom describing the activity hierarchy. For example:

\[
\text{travel\_by\_airplane} \sqsubseteq \text{travel\_by\_transport} \\
\text{travel\_by\_car} \sqsubseteq \text{travel\_by\_transport}
\]

We collect the set of possible activities based on POIs and the following classifications of activities. For example, Yellow Pages\(^5\).

\(^5\)http://www.yellowpages.ca/business/

### 4.1.2 Extended OSMonto

A POI denotes geo-referenced object such as a restaurant, a shop, or a lake using tags as described in Section 3.2. It activates a set of possible human activities that can be performed or hosted there or nearby. For example, on a HIGHWAY ROAD, TRAVELING BY TRANSPORTATION is the most usual activity.

In this study, we use the OSMonto ontology\(^6\) whose propose is giving a semantic interpretation of POIs. It mainly composes their identical information (e.g., university, school, restaurant, etc.) with a category in the form of a pair \(<\text{key}, \text{value}>\). Key is a category, and value is a POI. The key of a tag has the prefix "k_ ", and the value of a tag contains the prefix "v_ ". However, some tags can have the same values according to OSM tagging structure. For instance, the tag \(<\text{railway}, \text{train}>\), and the tag \(<\text{route}, \text{train}>\) have the same values. In this case, the values are a combination of the key and the value, as in \(k_\text{railway}v_\text{train}\).

We extended the ontology up to 493 POIs by adding new POIs that have recently appended to OSM (see, for instance \([13]\)) as well as the tags that frequently occur in the cell coverage map.

### 4.1.3 TimeOnto

Time of the day and day of the week are important factors for classifying the distribution of human activities over time. Each activity is associated with a time period which is represented in a more "qualitative" term than a fine grained time measurement. The qualitative terms fuzzify time periods. For example, early morning [6,8], mid morning [9,10] and late morning [10,11].

The Days and Times are hierarchically organized using the following axioms in Description Logic:

\[
\text{time} \sqsubseteq \text{day\_time} \\
\text{morning} \sqsubseteq \text{time} \\
\text{early\_morning} \sqsubseteq \text{morning} \\
\text{mid\_morning} \sqsubseteq \text{morning} \\
\text{workday} \sqsubseteq \text{day} \\
\text{holiday} \sqsubseteq \text{day} \\
\text{saturday} \sqsubseteq \text{holiday} \\
\text{sunday} \sqsubseteq \text{holiday}
\]

### 4.1.4 Relation between human activities and POIs

Based on the life style in a given region, we relate human activities to each POI. For example, a restaurant is annotated with eating and working activities. A university is annotated with studying, teaching, and working activities etc. So we associate human activities in ActOnto ontology to POIs in OSMonto ontology using the object property "what\_can\_be\_done\_eating". The assertion of this relationship \([11]\) is an existential restriction which is exemplified by the following axiom stating that for for every restaurant there is an activity of type eating, which is possible:

\[
\forall \text{restaurant} \exists ! \text{what\_can\_be\_done.eating}
\]

This way we associate around 220 activities with the POIs.

\(^6\)http://mayor2.dia.fi.upm.es/oeg-upm/files/hrmontology/hrmontology-RDF.zip

\(^7\)http://swat.cse.lehigh.edu/resources/onto/olympics.owl
We associate the times from TimeOnto ontology with the activities on workday early afternoons.

Figure 3: Example of classifier atomic formula for the human activities on workday early afternoons

### 4.1.5 Relation between human activities and days, times

We relate day-time periods to each activity based on the timetable of POIs and the characterization of POIs that activates an activity. We highlighted significant timetables of some POIs in the city of Trento, Italy where we perform the experimental evaluation:

- opening hours of shops, supermarkets [9am-12pm], [3pm-6pm], except Sunday
- opening hours of restaurants, coffee shops, cafeteria [8am-2pm], [6pm-10pm].
- opening hours of bars, clubs [8pm-3am].

We associate the times from TimeOnto ontology with the activities using object property "is_usually_done_during" and the days in TimeOnto to the activities using the object property "is_usually_done_on". The assertion of this relationship is the existential restriction, which is specified using the following example axiom in DL:

\[
\text{shopping} \sqsubseteq \text{activity} \sqcap \exists \text{is_usually_done_during.(morning \sqcap afternoon)} \sqcap \exists \text{is_usually_done_on.(workday \sqcap saturday)}
\]

This axiom should be read as "every shopping activity can be possibly performed in the morning and in the afternoon except on Sunday". This way, we associate all the activities to days and times.

### 4.1.6 Reasoning about human activities that occur at a certain day and time

A reasoning engine, Pellet\(^8\) reasoner, classifies a set of activities based on the time of a day. The example of a classifier atomic formula (see Figure 3) for the activities that could be activated on early workday afternoons is as follows:

\[
\text{workday_early_afternoon_activity} \equiv \text{activity} \sqcap \exists \text{is_usually_done_during.early_afternoon} \sqcap \exists \text{is_usually_done_on.workday}
\]

Around 42 classifier atomic formulas are written in HBOnto ontology.

8\(\text{http://clarkparsia.com/pellet/}

### 4.2 Stochastic model: a likelihood of human activities occurring

In the previous section, we described the Human Behavioral Ontology (HBOneto) model that captures the conceptual knowledge about the relation among a human activity, a POI, and a day, or a time. Making use of that and having the mobile network coverage map, we can estimate the likelihood of the human activities occurring in every area. This is done by considering the POIs in a area and the time/day. The stochastic model associates a likelihood to every human activity in a given area at a certain time of the day:

\[
P(a|l, t) = \frac{P(a, l, t)}{P(l, t)} \quad \text{(Conditional probability)}
\]

\[
P(a) * P(l|a) * P(t|a, l) \quad \frac{P(l, t)}{P(l, t)} \quad \text{(Multiplication Rule)}
\]

\[
P(a|l) * P(a, l|t) * P(l) * P(t) \quad \text{(Bayes Theorem)}
\]

\[P(a|l) * P(a, l|t) \quad (1, t \text{ independent variables})
\]

Moreover, the probability of the activities considering the nearby areas and given the time of a day is calculated as:

\[
P(a|l, l, r) = P(a, l, l|t) * P(a|l, r) \quad (l, r \text{ are not independent})
\]

### 4.2.1 Probability of human activities in a given area

\[- P(a|l)
\]

The POIs and the neighborhood POIs are the factors most affecting the likelihood of activity occurrence. Some POIs such as bus stops, small shops, ATM, etc., which frequently occur in several areas have a minor discriminative influence on the likelihood estimation than very characterizing POIs like an airport or a swimming pool. By borrowing the approach commonly used in the Information Retrieval theory, we adopted the Term frequency-Inverse Document Frequency (TF-IDF) function to estimate the importance of POIs in a given area based on the absolute occurrence. In this case, the Term Frequency (TF) factor measures the simple frequency of a POI in a given area, while the Inverse Document Frequency (IDF) factor gives an indication of the general discriminative power of a POI. A high IDF factor is associated with POIs that are rare while a low IDF factor is associated with POIs that are very common and thus are of low usefulness for distinguishing among different areas.

Since common a POI appears in many areas, the ratio inside the logarithm approaches 1, bringing the TF-IDF, IDF closer to 0. The POIs that occurring in a certain area have a higher weight, and the
POIs that occur everywhere have the lower weight. This way, after we assign a weight to each POI in a given area, we get:

\[ \text{tf} - \text{idf}(f,l) = \frac{N(f,l)}{\text{argmax}_w \{N(w,l) : w \in l\}} \cdot \log \frac{|L|}{|\{l \in L : f \in l\}|} \]  

(3)

1. \(f\) is a given POI; \(f \in F, F = \{\text{building, hospital, supermarket, ...}\}\)
2. \(l\) is a given location; \(l \in L, L = \{\text{cell1, cell2, cell3,...}\}\)
3. \(N(f,l)\) is the occurrence of the POI \(f\) appears in the location \(l\)
4. \(\text{argmax}_w \{N(w,l) : w \in l\}\) is the maximum occurrence of all the POIs in the location \(l\)
5. \(|L|\) is a number of all locations
6. \(|\{l \in L : f \in l\}|\) is a number of locations where the POI \(f\) appears

From the each POI, we retrieve a set of activities to each POI by performing the reasoning on the HBOnto ontology. The activity weight in a given location is a sum of the corresponding POI weight, as estimated as following equation 4.

\[ W(a,l) = \sum_{f \rightarrow a, f \in l} tf - idf(f,l) \]  

(4)

So we need to normalize the activity weights in a given location in order to estimate the activity probabilities conditioned by the location in Formula 5.

\[ P(a|l) = \frac{P(a,l)}{P(l)} \]  

(5)

1. \(P(a,l)\) is the probability of the activity \(a\) that occurs in the given location \(l; l \in L, L = \{\text{cell1, cell2, cell3,...}\}\)
2. \(P(l)\) is the probability of the activities that occur in the given location \(l\)

### 4.2.2 Probability of human activities at a certain day, time period - \(P(a,l|t)\)

As previously stated the day and time affect the likelihood of the activities. So we need to put a weight to the activities in order to rank the activities which could occur at the same time period. For instance, eating, working, studying, etc. In order to rank those activities, we consider the distribution of the activity occurrence over the time periods. As the initial step, we assume that if activity occurs given hour, the probability distribution of the activity is 1 or not 0. We use a Fuzzy Reasoning Model (FRM) (also referred to as approximate reasoning, see [7]) in order to estimate the importance weight of activities in different time periods. We assign membership degrees 0 (false) and 1 (true) to each element of the activities and to each element of the time periods, defined as a membership function (MF) which is represented in a curve. The graph visualization of the example fuzzy sets for some time periods and an activity is shown in Figure 2(b). In this figure, the x axis represents each hour in the universe of discourse(input space), whereas the y axis represents the degrees of membership in the [0,1] interval. The fuzzy membership values for the morning (curve in blue color) are defined between 6am and 11am. The fuzzy membership values for the having breakfast are defined between 6am and 10am. So, the importance weight of the having breakfast in the morning is the crisp intersection of their fuzzy membership values. The weight is estimated as following equation 6 in which \(\mu_{a_t}\) is a MF for the activity \(a\) and \(\mu_t\) is a MF for the time period \(t\).
The probability of an activity conditioned in a time period is estimated as following conditional probability 7.

\[
P(a, l| t) = \frac{P(a, l, t)}{P(t)}
\]

1. \(P(a, l, t)\) is the probability of the activity \(a\) of the location \(l\) in the given time period \(t; a \in A, A=\{\text{eating, working, studying,...}\}\)
2. \(P(t)\) is a the probability of the activities in the given time period \(t; t \in T, T=\{\text{early morning, mid morning, late morning, mid day,...late night}\}\)

4.2.3 Probability of human activities given nearby areas - \(P(a|l, r)\)
To avoid the spatial gap, we consider the nearby areas that located within the radius \(r\) of a circle from the given area \(l\). We estimate the maximum weight of the activities by Formula 8 and their likelihoods are estimated by Formula 5. The activity weight in nearby areas depends on the intersection weight \(\lambda_i\) defined in a range of values \([0,1]\) that measures the intersection between the nearby areas and the radius \(r\) of the circle.

\[
W(a,l,r) = \arg\max_{a} \{W(a,l_i)*\lambda_i\} : r \bigcap_{i=1}^{n} l_i
\]

1. \(\lambda_i\) is the intersection weight between the nearby area \(i\) and the radius \(r\) of the circle
2. \(\arg\max_{a} \{W(a,l_i)*\lambda_i\}\) is the maximum weight for the given activity \(a\), which occurs in the nearby areas that are located within the radius \(r\) of the circle;

5. EVALUATION
We designed an experimental setting for collecting data to validate our HRBModel. The goal of the experiment is to understand whether the human activities associated with a given area at a given time in our fuzzy model are correct by evaluating them with the feedback provided by a set of users. In this section, we introduce the design of the experimental application and the user feedback collection carried on the city of Trento, Italy. We then present the preliminary evaluation results.

5.1 Experimental Setting
In order to collect the list of human activities for any given area at different time slots, we created a web application to be evaluated by a set of selected users. Since the granularity of mobile network cell coverage can be different in the center of a city or remote area and also can be irregular shaped (see Section 3), we decided to create a customized spatial grid for the city in which human activities are uniformly distributed. The application offers the map of Trento city overlaid by the grid (see Figure 5) representing the mobile network cell coverage. First, the map has been divided into the grid of size 401 by 302, where each unit size of the grid is 50 m². Using the data coming from OSM, we populated the POIs at each area that contained, crossed, located on the border or intersected with that area. We discarded irrelevant POIs, which are not associated with any specific human activities. For example, bench, chimney, pipeline, garbage bins, etc. This results in a total of 195,548 POIs being populated.

Since the distribution of POIs in an area is very unbalanced (with few areas being very dense and most being very sparse), we reduced the number of the dense areas by grouping the areas so that each area contains a maximum of \(h\) POIs. We performed this operation using the quad tree data structure. A quad tree is a tree data structure used to partition a two dimensional space by recursively subdividing it into four quadrants or regions (square, rectangular, or of arbitrary shapes). So, we recursively partitioned the areas until they contained at most \(h\) POIs, and covered the minimum size of 50 m² area. For example, in a dense area (the center of the city), we kept the areas as it is size of 50 m², and in a sparse area (the mountains or nature), the areas were grouped into a larger size which contained at most \(h\) POIs. We chose the maximum threshold (\(h=38\)) number of POIs in each area that contained enough POIs to consider human activities there. This way we extracted 27,632 areas, with an average of 7.1 POI/area from the new partitions. For each area, we estimated the TF/IDF weight of all the activities. The relevancy of the TF/IDF weight and the occurrence of the POIs is depicted in Figure 6. The figure shows that while the number of the activity across all the areas is increased, the weight of the activity is reduced. For example, if the bus station is distributed in most of the areas, the weight is lower close to 0.

Our web application recommends the activities depending on the radius \(r\) of a circle from the given location and the given time as described in Equation 2. The \(r\) radius (the default one being 100m) is automatically chosen depending on the number of activities, at
### Table 1: Data collection of user feedback over different time periods

<table>
<thead>
<tr>
<th>Feedback on Workday</th>
<th>Feedback on Saturday</th>
<th>Feedback on Sunday</th>
</tr>
</thead>
<tbody>
<tr>
<td>morning 158, mid day 29, afternoon 59, evening 61, night 28</td>
<td>morning 19, mid day 5, afternoon 33, evening 19, night 8</td>
<td>morning 22, mid day 5, afternoon 25, evening 6, night 4</td>
</tr>
</tbody>
</table>

User feedback | Participant | Duration | Feedback clusters | Feedback in each cluster |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>481</td>
<td>32</td>
<td>1 week</td>
<td>5</td>
<td>Trento.Nord - 21, Downtown - 180, Povo - 151, Santa Chiara - 60, Trento.Sud - 67</td>
</tr>
</tbody>
</table>

Figure 6: Scatter plot of TF/IDF weights for the POIs in the city of Trento, Italy

Figure 7: Visualization of demonstration app, at Piazza Duomo on weekday mornings

Figure 8: Histogram of high level activities in the city of Trento, Italy

5.2 User feedback collection

We collected the user feedback through the web application described above for one week with 32 participants involved (see Table 1). It emerged that most of the user feedback comes from the areas of Trento-Povo and Trento-Downtown. The table shows the distribution of the user feedback over different time periods.

The feedback of every user is collected in a record containing: the latitude/longitude of the location selected on the map, the radius, the selected activity (among the top-5 proposed) or the one freely chosen, the semantic day and time (e.g., workday, saturday, etc.).

5.3 Outcome and preliminary evaluation

Given the collected feedback, we evaluated the HRBModel performance and present the preliminary results. We measured (i) the accuracy of the model and; (ii) the divergence between the activities likelihood extracted from our model and those coming from the collected users’ feedback. For the latter, we took into consideration only the areas with the highest number of feedback, Trento-Downtown and Trento-Povo. We also show the activity probability distribution divergence in a given time period compared to the activity probability distributions in our FRM.

To measure the accuracy of the approach, we consider the predic-
A total of 10 categories (shopping, eating, etc.) are around. The activity prediction to very generic high level activities has been increased to 80.2% when we increase the granularity of the activity correct (among the top-5) corresponds to 61.95%. The accuracy of the system performance in predicting the user activity regardless of its ranking position. Within the 481 item of users’ feedback collected, 341 activities have been correctly predicted, corresponding to an overall accuracy of 70.89%. The over-\[\sum_{t=1}^{T} P(a_i|t) \]

We also analyzed the divergence of the calculated activity probabilities comparing to the probability from the feedback in the areas with the highest amount of feedback: Trento-Downtown and Trento-Povo. The sample of results, Figure 8(a) and 8(b) describe the divergence of the probability of the activities that occur in those areas on weekday mornings and afternoons. There, we propagated the probability activities to their child activities. The activity indexing (x-axis) is different in each area. Y-axis is the probability associated with the activities. The graph shows that the activity probability activities from the user feedback can still follow the probability trend from our model. Those divergences enable us to improve the prediction task at run-time operation by learning the activity distribution divergence over time periods and locations. The FRM is a major input to the system for estimating the likelihood of human activities at particular time. Indeed people who live in the city of Trento usually do mountain sports more than any other activities, since the city has been surrounded by Alps mountains.

As shown in Figure 9, the main activities in the city are sporting, working, travelling by transportation and residential activity. As shown in Figure 9, the main activities in the city are sporting, working, travelling by transportation and residential activity. Indeed people who live in the city of Trento usually do mountain sports more than any other activities, since the city has been surrounded by Alps mountains.

In this paper, we introduced an ontological and stochastic HRBModel for the semantic interpretation of human behavior in a geographical area and at a given time that makes use of the geo-referenced knowledge available on the web. OSM and mobile phone cell coverage map. The accuracy of the model performance is good enough to predict and rank the possible human activities (top-k) in a certain place and at a certain time where the mobile network events occur. The design of an experimental evaluation and the preliminary results are also described and followed by the highlights for the prediction task improvement. The study aims at understanding the correlations between human behaviors and contexts, which allows us to associate the human activities with a particular network event (e.g., a sudden call amount peak) according to their likelihood. By leveraging this model, we will study the normal calling frequency patterns from the CDR that occur in given contexts, which allows us to successfully perform a set of predictive tasks such as the prediction of human activities under certain contextual conditions, the identification of behavioral patterns in various area profiles (e.g., business, shopping, or leisure areas etc.), and the characterization of exceptional events detected for anomalies in mobile network event data. The model can be used in a variety of analyses of many domains (e.g., smart cities, health-care, transportation, risk management, etc.) for their service creation and in combination with the following data sources: Twitter, gps, Foursquare, personal phone data, wireless data, etc.

In the future, in order to improve the activity prediction capability of HRBModel, we plan to enrich it with the other geo/time-referenced knowledge available on the web, such as weather forecasts, social events, business/daily activities of organizations, news...
events, statistical information about a region, and so on. We will employ online automatic extraction algorithms and tools to use those sources in real-time. The evaluation will involve a large number of participants, making use of the data with ground truth can be collected through crowd sourcing, user survey, and social networks (e.g., twitter, foursquare etc.) and so on.

7. REFERENCES


