What is this place? Inferring place categories through user patterns identification in geo-tagged tweets

Deborah Falcone
DIMES
University of Calabria, Italy
dfalcone@dimes.unical.it

Cecilia Mascolo
Computer Laboratory
University of Cambridge, UK
cecilia.mascolo@cl.cam.ac.uk

Carmela Comito
ICAR-CNR, Italy
ccomito@dimes.unical.it

Domenico Talia
DIMES
University of Calabria, Italy
talia@dimes.unical.it

Jon Crowcroft
Computer Laboratory
University of Cambridge, UK
jon.crowcroft@cl.cam.ac.uk

***Abstract*—Online social networks such as Facebook and Twitter have started allowing users to tag their posts with geographical coordinates collected through the GPS interface of users smartphones. While this information is quite useful and already indicative of user behavior, it also lacks some semantics about the type of place the user is (e.g., restaurant, museum, school) which would allow a better understanding of users' patterns. While some location based online social network services (e.g., Foursquare) allow users to tag the places they visit, this is not an automated process but one which requires the user help.**
In this paper we exploit the dynamics of human activity to associate categories to GPS coordinates of social network posts. We have collected geo-tagged tweets of a large city through Twitter. A supervised learning framework takes the tweets spatial-temporal features and determines human dynamics which we use to infer the place category. Our results over the data show that the prediction framework is able to accurately identify if a place is of a certain category given its user activity patterns. The average accuracy is about 70%, reaching the highest accuracy for work (90%) and educational places (80%). Moreover the framework identifies the category of a place, with an accuracy up to 66%, finding out where people eat and drink, go for entertainment, or work/study.

I. Introduction

The ability to associate spatial context to posts is becoming a popular feature of the most used online social networks. Facebook and Twitter exploit the GPS readings of users' phones to tag posts, photos and videos with geographical coordinates.

Some of the most recent location based online social networks such as Foursquare allow the user to explicitly indicate the place category he is at. While the category information is very rich and can enable more refined applications, this is a cumbersome manual process in which the user is voluntarily "checking into" a place. In most of the other social networking tools, a location is simply represented as latitude-longitude coordinates associated to a post automatically by the service. However knowing the semantics of the type of a place (home, office, museum) a user is at is potentially very useful. It could allow to infer users common interests, to improve activity prediction and ultimately mobile user recommendation and advertisement.

The aim of this work is to define and implement inference of location categories from geo-tagged posts in online social networks (specifically Twitter) of an urban area. The key aspect of the work is the extraction of spatial-temporal patterns from mobile users tweets. Through these, we have built a framework to infer the category of the visited places among a finite set of alternatives. We address the problem as a supervised classification task.

As a proof of concept we perform a fine grain analysis of a large city (i.e., London) geo-tagged dataset obtained from Twitter. The proposed methodology consists of various phases. Due to the variable accuracy of GPS, the preliminary step is to *cluster the locations* of the geo-tagged tweets so that each place is identified by a single pair of geographic coordinates. We then assume that for some locations we can associate each place to the most-likely category extracted from a database of categories and coordinate associations (namely a *Foursquare database*) that represents the ground truth labels. This will be the class attribute exploited by the supervised learning algorithm which is able to automatically label places. The next step is the extraction of *spatial-temporal information* for each labeled place: for each user we compute a set of daily snapshots extracting the visited places and the movements among them, the duration of the visits at each location, mining also when and how many times this place is visited during the day and during the week. Finally, we aggregate this information for each place visited in the considered period. Based on these patterns we define a set of machine learning features that together with the ground truth categories of Foursquare are the input of our machine learning algorithms able to classify unlabeled places given their similarity to other

places of which we know the categories.

The novelty of our work consists in identifying place semantics based purely on spatial-temporal features such as stay duration, time of day, place popularity, extracted from irregular social posts.

The contributions of this paper can be summarized as follows.

- We defined a method to extract spatial-temporal patterns from geo-tweets which exploits a number of features specific to the places, the duration of the stay, the time of day and of week of the typical stay, the number of visitors and the regularity of their behavior in the place.
- We designed two classification tasks for place labeling:
  - The **Binary problem** aims to infer whether a place belongs to a certain category. Results demonstrate that we can correctly answer with an average accuracy of about 70%, reaching the highest accuracy for work and educational places (90% and 80%, respectively). This classification can be useful for instance to decide if a specific ad which is related to food promotions is appropriate for a location (i.e., only if the location is a restaurant or a bar).
  - The **Profiling problem** allows to infer the category of a place among a set of categories. In this study, we consider three different category sets that typically characterize people's everyday life. Our results show that with an accuracy of 66% we can identify if a place is one in which people eat and drink, or it is a leisure place, or a place for work/study.

The rest of the paper is organized as follows. Section 2 overviews related work. Section 3 motivates our work, while in Section 4 we describe the Twitter dataset. Section 5 explains our methodology, introducing the key aspects of our approach. In Section 6 we show the results of our evaluation. Finally, Section 7 concludes the paper.

II. Related work

Social networks access from mobile devices has increased dramatically in the last few years, generating huge amount of geo-tagged social data as often GPS readings are associated to social network posts and general activity by services. Semantic place labeling has become a popular research direction given the importance of associating context to geographical coordinates. Various approaches for semantic place labeling have been proposed recently.

In the work by Liao et al. [9] the authors used features including the time of visits and the presence of bus stops, restaurants, and other points of interests to automatically label places. The main innovation is the introduction of a hierarchical conditional random field (CRF) that aids inference accuracy by exploiting the temporal sequence of place visits, e.g., *works* often follows *home*. Their system was tested on GPS trajectories of only four people. Chen et al. [4] followed a similar approach but they processed label sequences with a hidden Markov model rather than a CRF. Our approach attempts to automatically label places based on a mapping between features describing a visit to a place and the place label.

In [8] the authors developed a classifier called Placer that classifies locations into different label categories based on the timing of visits to that place, nearby businesses, and simple demographics of the user. Our work shares the same aim as [8] of inferring location category by exploiting the timing visits to a place. However, this is the only common feature. We aim to classify locations by exploiting the tweets whereas [8] uses as data sources two different diary surveys conducted in the U.S and only one of which includes latitude/longitude data. Another important difference between the two approaches is that while in our case we have to identify and extract features by analyzing the tweets, in [8] the statistics used as features are already available from the surveys not requiring, thus, any analysis and processing of data as it is instead necessary in our case. Furthermore, the features used in [8] are different from the ones we used. In fact, [8] uses demographic data like the age and the gender of users together with nearby business and point of interests (POI) as features for the classifier. Similarly to us, they use temporal features but at a much coarser grain level: the only commonality in this case concerns the feature that accounts for the duration of the visit.

To the best of our knowledge the only work that similar to ours identifies place category from social network data is the one by Ye et al. [12]. In [12] authors derived eight place label categories from Whrrl, a location-based social network. They used a support vector machine on features such as check-in frequency and time of day to label over 53,000 places from almost 6,000 users. They identified and extracted features, however not temporal features such as the duration of visits or other timing statistics which we will use.

The work in [13], like ours and [12], [8], uses a machine learning approach to automatically build a place classifier, defining the semantic place annotation as a classification problem. In particular, this work is very similar to ours in the choice of the features. It is the first work that uses as features temporal values taking into account the duration of the visit in a given place. However, also with respect to this work, our approach presents elements of novelty. First, the work in [13] is for GPS traces of mobile devices whereas we deal with the more unpredictable and irregular data of the Twitter social network. Second [13] uses also movement-related features by exploiting the accelerometer of the mobile devices. Third, we use more elaborated and effective features compared to the one defined in [13].

Other approaches exploit geo-tagged Twitter data together with the semantic content of tweets to model crowd mobility [11], [5], [7]. In particular, Kinsella et al. [7] create geographic language models at varying levels of granularity (from zip codes to countries) to predict the location of the tweet based on linguistic content changes.

III. Problem motivation and formulation

The problem of automatically associating semantics to geographical location is one which has attracted researchers

attention as it enables a variety of rich applications ranging from recommendation, advertisement and better space planning. For instance, knowing if a place is of a certain category could help recommending targeted places, e.g., tourists could be recommended leisure places or monuments specifically. While, knowing the category of a place a user is in could be useful for automatically inferring activities of people, e.g., if a user is in a work place, he is likely working. In this section we provide a formulation for our location category inference task. Given a location in a city expressed exclusively as geographical coordinates, our aim is to label it on the basis of the spatial-temporal patterns exhibited by users in that location. More precisely we formulate this problem in terms of a specific social network, Twitter, and its geo-localized tweets, although the general formulation is applicable to any other social network with geo-tagged posts. Given a set of geo-tagged tweets TW, we extract the set of locations L visited from Twitter users U.

A tweet $\mathrm{tw} \in \mathrm{TW}$ is characterized by: the timestamp (t), that is the time at which it has been posted; and the location (l) from where it has been posted. Each location $\mathrm{l} \in \mathrm{L}$ is represented by a triple: $\mathrm{l} =< (\mathrm{lat}, \mathrm{lon}), [\mathrm{p_l}], 4\mathrm{sq_l} >$ where: $(\mathrm{lat}, \mathrm{lon})$ is the pair of geographic coordinates that identify the location; $[\mathrm{p_l}]$ is an array containing its properties; and $4\mathrm{sq_l}$ is the category of the most-likely Foursquare venue that can be associated to l. We use Foursquare purely as a source of a ground truth database of labels and any other database would have fulfilled the same purpose. The properties vector $[\mathrm{p_l}]$ of a location l contains: $\mathrm{U_l}$, the set of people who visited l; $\mathrm{TW_l}$, the set of tweets posted in l; $\mathrm{V_l}$, the set of the visits of l; $\mathrm{tIn_l}$, the total time that users spend in l; and $\mathrm{nD_l}$, the total number of days in which l is visited.

A visit $\mathrm{v} \in \mathrm{V}$ is characterized by: u, the user who visits the location l; $(\mathrm{tw_{first}}, \ldots, \mathrm{tw_{last}})$, a sequence of tweets that u posts in l before moving to another place; and $\Delta\mathrm{v} = \mathrm{tw_{last}} - \mathrm{tw_{first}}$, the duration of the visit equals to the difference between the timestamp of the last and the first tweet of the visit. In accordance to these properties, we define a set of learning features used as input vector for our machine learning classifiers. Formally, we can define the supervised classification problem of inferring the category of a place in the following way:

Given a location l *and its spatial-temporal patterns, we aim to infer its category* c *among a finite set of categories* C.

Before giving the details of the classification task and the features used, we will now describe data on which we conduct our study.

IV. Twitter dataset and temporal patterns

The geo-located data mined in this work is a dataset of tweets tagged with GPS location within the boundaries of the city of London, one of the top three cities by number of tweets[1]. Numerically speaking, we consider a Twitter dataset of 7,424,112 tweets issued by 292,195 mobile users in 6,098,148 distinct locations, during a period of six month started in June 2013 and ended in November 2013. We built a multi-threaded crawler to access the Twitter Streaming API. The crawler collects the tweets filtered by location and processes the results to obtain a dataset in which each entry is a tweet that includes the ID of the user who created the tweet, the timestamp and the GPS coordinates of the tweet. The dataset represents a sequence of daily snapshots, with an average number of tweets per day greater than 40,000. The data analysis reveals that the behavior of users is very heterogeneous: note the long tail of the probability distribution functions (PDF) both of the number of tweets and of the time interval that elapses between successive users' tweets. Figure 1 shows the PDF of the number of tweets per user in a month. Even if the volume of tweets per month is very high, most of the users (78%) post less than 10 tweets per month. This may depend on the fact that many users are tourists and then occasionally visit the city.

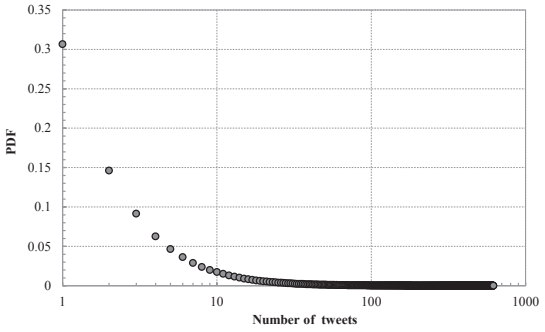


Fig. 1. Probability Distribution Function of number of monthly tweets per user.

A similar pattern arises considering the time elapsed between successive tweets. Figure 2 shows that 60% of tweets are posted with an inter-tweet intervals less that one hour. In particular about 40% of these are posted with high frequency, i.e., with an inter time of 10 minutes. Moreover, the graph shows that only 28% of tweets are posted with a frequency greater than 3 hours. A similar trend is observed in [3], with about 46% of intervals between tweets greater than or equal to one hour.

We also observe the tweets frequency during the course of the week. Figure 3 shows that the tweets rate for each day of the week has a periodic behavior. Days exhibit a peak in the evening and a dip at night time. The figure also highlights some differences between week days and the weekend. In particular, in weekends the volume of tweets is higher, mainly during the morning and there is a peak at lunchtime. This is more evident on Saturdays.

These patterns seem to mirror user behavior: for instance, during a week day, a user might spend morning and afternoon at the workplace, taking a lunch break in a restaurant, while in the evening he might go to the gym, to the cinema or stay at home. In order to exploit these temporal patterns for our classification task we divide the day into six different time

[1]http://semiocast.com/

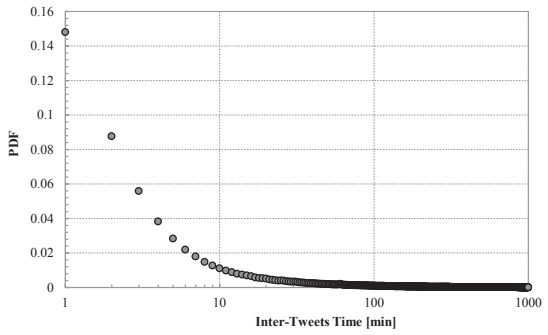


Fig. 2. Probability Distribution Function of the time elapsed between consecutive tweets.

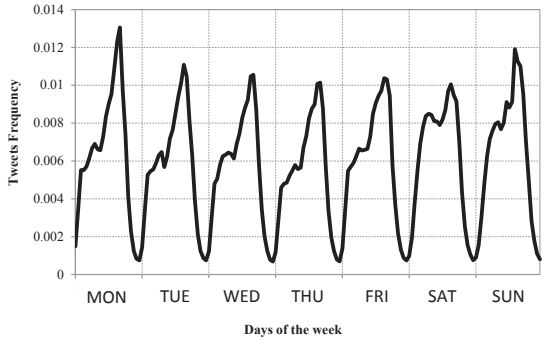


Fig. 3. Tweets frequency during a week.

slots, as shown in Figure 4. The time slots, are formally specified by the following:

Def 1. TS is a finite set of time slots with $|\mathrm{TS}| = 6$. Each $\mathrm{ts} \in \mathrm{TS}$ is a time-object of varying time duration belonging to a day. $\mathrm{TS} = \{\mathrm{N}, \mathrm{EM}, \mathrm{M}, \mathrm{A}, \mathrm{EE}, \mathrm{E}\}$ where:
$\mathrm{N} = \mathrm{Night}[12:00\mathrm{am} - 05:59\mathrm{am}]$;
$\mathrm{EM} = \mathrm{EarlyMorning}[06:00\mathrm{am} - 09:59\mathrm{am}]$;
$\mathrm{M} = \mathrm{Morning}[10:00\mathrm{am} - 01:59\mathrm{pm}]$;
$\mathrm{A} = \mathrm{Afternoon}[02:00 - 05:59\mathrm{pm}]$;
$\mathrm{EE} = \mathrm{EarlyEvening}[06:00\mathrm{pm} - 08:59\mathrm{pm}]$;
$\mathrm{E} = \mathrm{Evening}[09:00\mathrm{pm} - 11:59\mathrm{pm}]$.

On the basis of this definition, we specify a mapping function $\mathrm{TS}(\mathrm{tw})$ to associate the corresponding time slot to the timestamp of a tweet.

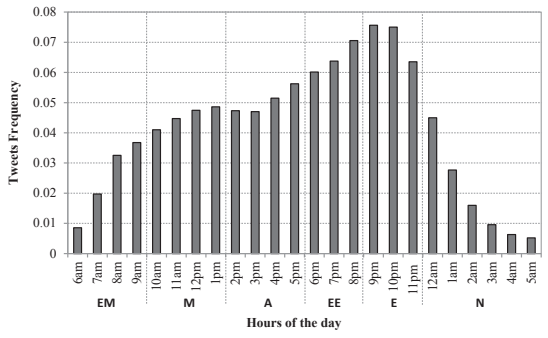


Fig. 4. Temporal evolution of tweets frequency during the daily time slots.

V. Location categorization methodology

Since GPS precision is 10 meters on average, a specific semantic location in the city might be represented by slightly different GPS coordinates. Before we dive into the description of our supervised learning place classification we show how we have clustered coordinates into places and associate them to Foursquare categories when available.

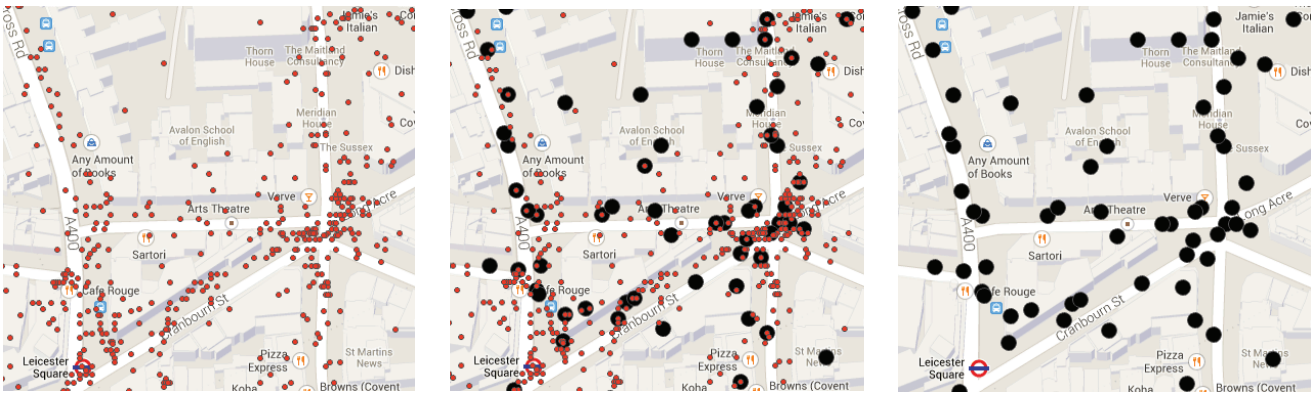
A. Location Clustering

We have used a clustering algorithm that receives as input a set of geographic coordinates L extracted from a set of geo-tagged tweets TW. Each location $\mathrm{l} \in \mathrm{L}$ may be represented by one or more pairs $(\mathrm{lat}, \mathrm{lon})$. What we want to achieve by applying the clustering algorithm is that each $\mathrm{l} \in \mathrm{L}$ is uniquely identified by a couple of geographical coordinates. To this aim we use OPTICS [2], an algorithm for finding density-based clusters in spatial data (although other algorithm could also have fulfilled the propose). We have exploited the implementation provided by ELKI [1], an open source data mining software focused on unsupervised methods in cluster analysis and outlier detection. The output has a hierarchical structure, therefore we define a custom method to extract a flat clustering from a cluster tree. Our algorithm performs a top-down visit of the hierarchy; in particular it visits the tree from root to leaves, stopping when it found a locally optimal cluster. A cluster is locally optimal if all its elements are less than 15 meters from its centroid. The result is a set of clusters C. Each cluster $\mathrm{c} \in \mathrm{C}$ groups a subset of L that respect space distance constraints, and it is identifies by its centroid $\dot{\mathrm{c}}$. The evolution of the clustering algorithm is shown in Figure 5. This approach is able to identify specific venues in a city.

B. Category Association

Once the clusters are identified we want to associate them with the most-likely place type that represents the ground truth semantic place label (e.g., restaurants, schools, gyms). The database of labeled places will be used by our classification algorithm to aid the geo-tweets classification. For the evaluation of this work we retrieve a mapping of each geo-tweet to a semantic location. We use these mapped locations for training our framework and to validate the prediction (more in the evaluation section). The automatic labeling of places of an urban area, is the objective of our work. To this aim we use a Foursquare database of London, retrieved by the Foursquare API as ground truth labels. This database contains 39,304 Foursquare venues. For each venue we consider the pair $(\mathrm{lat}, \mathrm{lon})$ that identifies the location and the Foursquare category *4sq*. The main parameter of our mapping algorithm is the maximum mapping distance δ, that is the maximum allowable distance between a location to be mapped and a Foursquare location. Following different experiments we chose $\delta = 25$ meters. The output is a set of labeled clusters.

We refer to the eight top categories of Foursquare, that are *Professional&OtherPlaces*, *College&University*, *Nightlife&Spot*, *Food*, *Shop&Service*, *Arts&Entertainment*,



(a) (b) (c)

Fig. 5. Evolution of Clustering Algorithm: the geo-locations of tweets (a) are grouped by the algorithm so that each cluster respects spatial distance constraints (b). Each cluster of geo-locations is represented by its centroid (c).

Outdoors&Recreation, and *Travel&Transport*.

Many locations cannot be mapped with a Foursquare venue (mainly streets). Due to this, it is not possible to assign a category to all the clusters: in particular 58,119 clusters could not be associated to a category. As a result, we worked on 33,680 clusters. The distribution of places among the categories is depicted in Figure 6. As expected, the category that includes the largest number of places is *Food*. *Shop&Service*, *Professional&OtherPlaces*, *Travel&Transport* and *Nightlife&Spot* have a substantial set of places, while the other categories include less than 10% of the total places.

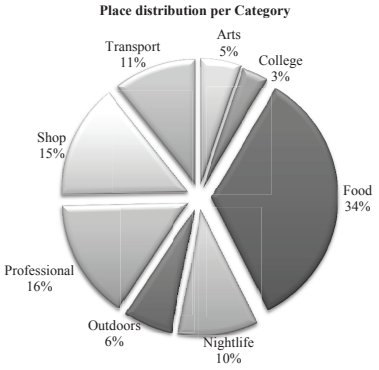


Fig. 6. Places distribution among the eight Foursquare categories.

C. Spatial - Temporal Patterns of Places

In this section we discuss one of the key point in our analytical approach: the extraction of spatial-temporal patterns for each place. On the basis of this information we will define the set of machine learning features.

First of all we introduce the concept of *consecutive tweets*, based on the time slots defined above.

Def 2. Two tweets tw_1 and tw_2, temporally ordered, are consecutive iff:

$$TS(tw_1) = TS(tw_2) \vee TS(tw_2) = succ(TS(tw_1)) \wedge \quad if \quad TS(tw_1) \neq N, \quad \delta(tw_1, tw_2) < 3h \tag{1}$$

where $succ(\mathrm{TS}(\mathrm{tw}_1))$ is the successive time slots of tw_1 and $\delta(\mathrm{tw}_1, \mathrm{tw}_2)$ is the temporal distance between the tweets.

This constraint guarantees that the time elapsed between tw_1 and tw_2 is not too long. If they are posted in the same time slot, the maximum temporal distance allowed is equal to the length of the time slot, that is limited. Moreover, if tw_1 and tw_2 are posted in successive time slots, the maximum distance allowed is three hours. If the time slot of the first tweet is night (N), we relaxed the constraint and the second tweet can be posted during the entire next slot, that is the early morning (EM). This choice is based on the observation that most people sleep during night, so we could not have tweets during that time period. Therefore, it is reasonable to think that in the early morning people are still in the place where they slept, even if there are no tweets to substantiate that.

In a day d a user u might visit one or more locations. For each user we compute his daily trajectories. A *daily trajectory* is formalized as follows:

Def 3. A daily trajectory, $\mathrm{DT}_{\mathrm{u,d}}$, is a sequence of visits at different locations $(\mathrm{v}_{\mathrm{l}_0}, \mathrm{v}_{\mathrm{l}_1}, \ldots, \mathrm{v}_{\mathrm{l}_n})$ by a user u during a day d:

$$\mathrm{DT}_{\mathrm{u,d}} = \mathrm{v}_{\mathrm{l}_0} \longrightarrow \mathrm{v}_{\mathrm{l}_1} \longrightarrow \ldots \longrightarrow \mathrm{v}_{\mathrm{l}_n}$$

When two consecutive tweets tw_1 and tw_2 of a user u are posted from two different locations l_1 and l_2, obviously the user has moved from one place to another. We assume that during the time interval between tw_1 and tw_2, referred to as *movement time*, $\mathrm{m}_{\mathrm{l}_1,\mathrm{l}_2}$, the user u was in l_1 (of course this is an assumption which does not factor in the transit time: we leave this improvement for future work).

If tw_1 and tw_2 are not consecutive (in other words, the time between them is too long) nothing can be assumed of the whereabouts of the user in that time frame, and we do not take into account the time interval between tw_1 and tw_2. The same applies when two not consecutive tweets are made in the same location l. In this case we consider the tweets as referring to two different visits to l.

A daily trajectory therefore contains distinct visits to locations;

each visit to a location l, namely $\mathrm{v_l}$, is characterized by the sequence of consecutive tweets.

For each visited location l in $\mathrm{DT_{u,d}}$ we compute the daily time spent in it by the user u during the day d, indicated with $\mathrm{tIn_l^{u,d}}$, and formalized as follows:

$$\mathrm{tIn_l^{u,d}} = \sum_{\mathrm{v_l} \in \mathrm{DT_{u,d}}} \Delta \mathrm{v_l} + \sum_{\mathrm{v_{l'}} \in \mathrm{DT_{u,d}}} (\mathrm{m_{l',l}}) \qquad (2)$$

Where:

- $\sum_{\mathrm{v_l} \in \mathrm{DT_{u,d}}} \Delta \mathrm{v_l}$ is the overall visits duration,
- $\sum_{\mathrm{v_{l'}} \in \mathrm{DT_{u,d}}} (\mathrm{m_{l',l}})$ is the overall time for movements started from the location l by u in the day d.

Notice that, only the time intervals between two consecutive tweets, will be considered in the calculation of the visit duration.

To clarify the above, we show an example of a daily trajectory of user u:

$$\mathrm{DT_{u,d}} = \mathrm{v_{l_1}}^{(7,N)} \longrightarrow \mathrm{v_{l_2}}^{(4,EM)(5,M)} \longrightarrow \mathrm{v_{l_2}}^{(8,E)} \longrightarrow \mathrm{v_{l_3}}^{(3,E)}$$

For each visited location l is shown a list of couples; each couple is the number of consecutive tweets posted in l in the timeslot TS. In the example, the user visits 3 different locations. At night time (N) u posts 7 consecutive tweets in $\mathrm{l_1}$, and moves during the early morning (EM) in $\mathrm{l_2}$. The time that the user spends in $\mathrm{l_1}$ is the period from the first tweet in N in $\mathrm{l_1}$ and the first tweet in EM in $\mathrm{l_2}$. u posts 4 tweets during EM and 5 tweets during M in $\mathrm{l_2}$, and all the tweets are consecutive. After that, his next tweet is recorded too long after in the evening (E). This means that the last tweet posted in M and the first posted in E are not consecutive. We cannot say that the user spends also the afternoon (A) and the early evening (EE) in $\mathrm{l_2}$, so we consider two different visit to the location. From $\mathrm{l_2}$ the user moves to $\mathrm{l_3}$ during the evening (E). The time that the user spends at $\mathrm{l_2}$ is composed of the sum of the time spent in it during the two visits, considering the consecutive tweets, plus the time elapsed between the last tweet in E in $\mathrm{l_2}$ and the first tweet posted during E in $\mathrm{l_3}$. Since we do not know where the user is after $\mathrm{l_3}$, the time of the visit in $\mathrm{l_3}$ is composed just of the difference between the last and the first tweet posted in it.

It is interesting to observe the trends of the temporal and spatial distance among the movements. Figure 7 shows the Complementary Cumulative Distribution Function (CCDF) of the temporal distance between the sequential tweets. The log-log plot presents a power-law functional form with a long tail indicating a wide variety of temporal measures. In fact, about 38% of movements occur frequently, 30% of them occur with an inter-time distance that varies between 10 minutes and 60 minutes, and the rest of the movements have a temporal distance greater than 1 hour. This information is important for our work: this means that there are only a few cases that we cannot consider due to a temporal distance being too long.

Likewise, the distribution of the space distances that users travel when moving from one location to another is a power-low, as shown in Figure 8. In this case the CCDF highlights

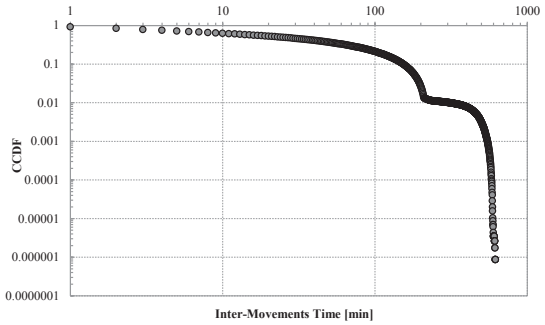


Fig. 7. Complementary Cumulative Distribution Function of the inter-movements time.

that most of users tend to move for short distances. In particular, 48% of movements cover a radius greater than 100 meters but within a distance of 1 km, and only 4% of movements affects a distance of more than 10 km.

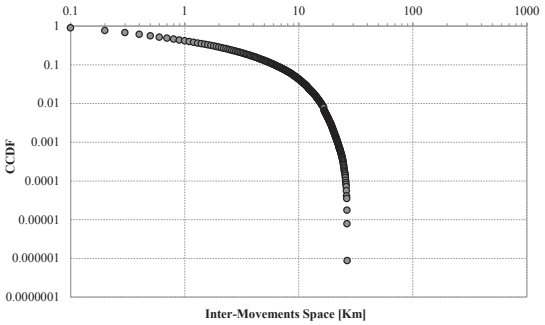


Fig. 8. Complementary Cumulative Distribution Function of the inter-movements space.

D. Classification Features

Now that we have i) labeled the locations through the database matching and ii) analyzed user behavior patterns at locations, we aim to exploit those in our classifier to allow the labeling of unknown locations automatically. Our classifier must infer the category of a place based on the spatial-temporal patterns. In this section we derive a set of machine learning features that together with the ground truth places category will be the input for the supervised classification algorithms. The features answer the following questions: (1) *how many people* visit the place? (2) *how long*/ (3) *when*/ (4) *how is* the place visited? The features identified are listed below.

Number of visitors. Knowing the number of people who visit a place is indicative of its popularity. The most visited places in an urban area are the public places such as squares and parks (*Outdoors&Recreation*), places of entertainment like museums and theaters (*Arts&Entertainment*), and certainly, all transport hubs such as subways, railway stations (*Travel&Transport*). The number of visitors of a location l, is the cardinality of $\mathrm{U_l}$, defined above, that includes the distinct users who have posted at least one tweet while at the location.

$$\mathrm{nVisitors_l} = |\mathrm{U_l}| \qquad (3)$$

Daily User Stay. One of the most important aspects to capture is the daily time that users spend in a place. People spend on average 60% to 65% of their time at home and between 20% and 25% at work or college [10]. In accordance to this, we expect that the time that users spend in a work place (*Professional&OtherPlaces*) and in an educational place (*College&University*) is longer than the time elapsed in other places like bar or restaurant (*Food*), boutique or Internet cafe (*Shop&Service*), train or tube station (*Travel&Transport*). For each user u and location l we measure the average daily time spent in l by u, ($\mathrm{dTime_{l,u}}$). We compute the average daily time among all the users that visit the location, and is defined as:

$$\begin{aligned} \forall \mathrm{u} \quad \mathrm{dTime_{l,u}} &= \sum_{\mathrm{d}} \mathrm{tIn_l^{u,d}} / \mathrm{nD_{l,u}} \\ \mathrm{DailyUserStay_l} &= \frac{\sum_{\mathrm{u}} \mathrm{dTime_{l,u}}}{|\mathrm{U_l}|} \end{aligned} \tag{4}$$

Where $\mathrm{tIn_l^{u,d}}$ is the time that a user u spent in l on day d, as expressed in equation 2, and $\mathrm{nD_{l,u}}$ is the number of days in which u visits l.

Short and Long Time Visit. We define a pair of machine learning attributes that provide information on the duration of the single visits. Our aim is to separate short from long visits. As an example, a user could visit the same bar several times within a day (maybe the bar near his work place). All the visits are very short except one during the lunch break. The daily time will be the sum of all the time periods spent in that place, which can be quite long in relation to the time spent for having lunch, even if most of the visits to this place are short. Furthermore, a place could be visited by people with different profiles. For instance, a post office could be visited by a lot of short stay customers, and by a small number of employees that spend all the working day there. We define the short visits feature as:

$$\begin{aligned} shortVisits_l &= |v \in V_l : \Delta v < 30min| \\ ShortTimeVisit_l &= \frac{shortVisits_l}{|V_l|} \end{aligned} \tag{5}$$

Where $\Delta \mathrm{v}$ is the time length of the visit, and $|\mathrm{V_l}|$ is the total number of visits in l as defined above. Similarly, the long visits feature is described as:

$$\begin{aligned} longVisits_l &= |v \in V_l : \Delta v > 3h| \\ LongTimeVisit_l &= \frac{longVisits_l}{|V_l|} \end{aligned} \tag{6}$$

Night Location. This feature is proposed to separate the locations visited mostly at night time from those visited during the day. This is formally achieved by measuring the ratio between the night time ($\mathrm{nigthlyTime_l}$) and the total time spent at a location. It is expected that this value will be greater for nightclubs and pubs (*Nightlife&Spot*), or places such as bowling or theaters (*Arts&Entertainment*), rather than places belonging to categories like *Professional&OtherPlaces*, *College&University*, and *Shop&Service* categories.

$$\begin{aligned} \mathrm{nigthlyTime_l} &= [\mathrm{t_0}, \mathrm{t_1}] : \mathrm{t_0} \in \mathrm{EE} \wedge \mathrm{t_1} \in \mathrm{N} \\ \mathrm{NightlyLocation_l} &= \frac{\mathrm{nightlyTime_l}}{\mathrm{tIn_l}} \end{aligned} \tag{7}$$

Weekend Location. Likewise, we aim to partition the locations visited mostly during the weekend from those visited during weekdays. We expect that the time spent in a location during the weekend ($\mathrm{weekendTime_l}$) will be higher for categories of places such as *Nightlife&Spot*, *Outdoors&Recreation*, *Arts&Entertainment*. Conversely, weekdays are mainly working days: the most popular places will be those of categories *Professional&OtherPlaces* and *College&University*.

$$\begin{aligned} \mathrm{weekendTime_l} &= [\mathrm{t_0}, \mathrm{t_1}] : \mathrm{t_0} \in \mathrm{Sat} \wedge \mathrm{t_1} \in \mathrm{Sun} \\ \mathrm{WeekendLocation_l} &= \frac{\mathrm{weekendTime_l}}{\mathrm{tIn_l}} \end{aligned} \tag{8}$$

These metrics allow us to keep into account: the number of visitors at each location, how long the place is visited and when the visits occur. We now formalize how a place is visited. In this sense we must consider whether the location is a transient or a steady place, i.e., if users visit a location at usual times or they transit in it without any regularity. For this purpose we use the Shannon Entropy:

$$H(X_l) = -\sum_{u=1}^{n} p(x_u) logp(x_u), \quad where \;\; p(x_u) = f(l,u) \tag{9}$$

to measure the uncertainty of three variables:

- $\mathrm{H(T_l)}$, tweets in the location l;
- $\mathrm{H(F_l)}$, frequency of the tweets in the location l;
- $\mathrm{H(D_l)}$, days in the location l.

We now describe these in detail.

Tweets Entropy. This feature tells whether users tend to tweet regularly in a location l, describing the distribution of its tweets across the users. For the feature $\mathrm{H(T_l)}$, $\mathrm{f(l,u)}$ is the user's proportion of tweets at the location l and is defined as:

$$f(l,u) = \frac{|TW_{l,u}|}{|TW_l|} \tag{10}$$

Where $\mathrm{TW_{l,u}}$ is the set of tweets posted in l by user u, while $\mathrm{TW_l}$ is the whole set of tweets posted from that location. We expect a small entropy in the *Professional&OtherPlaces* or *College&University* places in which people tend to have more stable and periodic behavior. In contrast, a higher entropy value implies that many users tweet from l, but they have very few tweets. This user behavior is typical in *Arts&Entertainment* or *Nightlife&Spot* places.

Frequency Tweet Entropy. Another important aspect to be highlighted is the tendency of users to tweet frequently (or less frequently) from a place. It is expected that in the workplace users do not often post messages (*Professional&OtherPlaces* or *College&University*). On the other hand, during leisure time, while sightseeing or visiting a museum, the frequency of tweets could be greater: someone could post photos or information about the visited place (*Arts&Entertainment*,

Outdoors&Recreation, *Shop&Service* or *Travel&Transport*). The feature $\mathrm{H(F_l)}$ is expressed formally by defining $\mathrm{f(l,u)}$ as the user's proportion of frequent tweets from location l:

$$f(l,u) = \frac{|\tau_{l,u} : \tau_{l,u} < 15min|}{|\tau_l|} \quad (11)$$

Where $\tau_{\mathrm{l,u}}$ is the set of short time intervals (i.e., shorter than 15 minutes) elapsed between two tweets posted from l by user u, while τ_{l} is the set of all the inter-tweets times at l.

Daily Visit Entropy. This feature is correlated to tweets entropy, but it further emphasizes users regularity for location l. The feature captures whether users visit a location in a periodic way, returning on different days. $\mathrm{H(D_l)}$ is defined as:

$$f(l,u) = \frac{nD_{l,u}}{nD_l} \quad (12)$$

Where $\mathrm{nD_{l,u}}$ is the number of days in which a user u visits the location l, while $\mathrm{nD_l}$ is the total number of days in which l is visited, as defined above. For instance, in a work or educational place (*Professional&OtherPlaces* or *College&University*), users might exhibit periodic behavior (e.g., five days a week), whereas leisure or food places may attract many opportunistic visits.

In the following we show how a classifier can be trained over the features extracted to associate place categories to geographical coordinates automatically.

VI. Using features to associate categories to places

The features defined in the previous section are used to construct classification tasks, using a supervised approach. Foursquare categories are fed to the classifier, together with the feature vector. We have used six classification algorithms, each one based on a different data mining classification technique: J48, Decision Table, Multilayered Perceptron, Bayesian Network, K* and LogitBoost. The algorithms are those included in the collection of Weka, a popular open source data mining toolkit [6].

In order to estimate the accuracy of our prediction algorithms, we used the 10-fold cross-validation as model validation technique and the set of metrics typically used in classification problems: precision (*P*), recall (*R*), and f-measure (*FM*). The overall task consists of assigning a decision class label (the category) to a set of unclassified locations, described by the defined set of features. *Our purpose is to infer the category of locations in a city knowing a set of categories* C *that may be associated with those locations*.

Binary problem. In this formulation of the problem the cardinality of C is one. This means that we are interested in finding if a place as isolated by the GPS coordinates of a tweet (and then clustered as described in Section V) belongs to a certain category or not. Formally, this problem results in a binary classification problem:

Given $\mathrm{c} \in \mathrm{C}$*, is* c *the category of a location or not?*

For each category we construct an annotated training dataset, in which we have two groups of instances. One group belongs to the category c and all its elements are labeled with *Yes*, while the other group is a random sample of the population. In the latter the instances are labeled with *No*. Figure 9 shows the average accuracy achieved for each Foursquare category by the six classifiers. The highest average accuracy is obtained for categories *Professional&OtherPlaces* and *College&University*. We can label the work places with an accuracy of 90%. *College&University* has average accuracy at 81%, followed by *Outdoors&Recreation* averaging 67%. *Arts&Entertainment* and *Travel&Transport* have moderate average accuracy at 63%. For the classes *Food*, *Nightlife&Spot* and *Shop&Service*, we observe that the average accuracy is slightly less than 60%. The latter value may be justified from the diversity of places that each of these classes represents. For instance, *Food* includes bars and restaurants. Not only the time spent in a bar is generally shorter than the time spent in a restaurant, but also, people tend to visit a bar at different times during a day, while a restaurant is visited mainly at lunch time and dinner time. For the purpose of recommendation, this level of category clustering however seems quite sufficient. On the contrary, work places tend to be visited according to a unique temporal pattern.

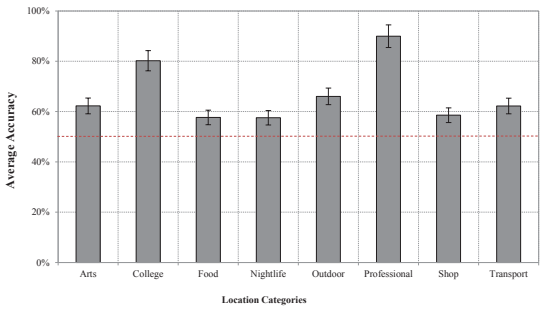


Fig. 9. Average Accuracy of the eight location categories inferred by the six classifiers.

In Table I we show the values of the validation metrics precision (*P*), recall (*R*), and f-measure (*FM*) of the most performing classification algorithms for each category class. As evidenced by the values in the table, the accuracy is very high for the class *Professional&OtherPlaces*, with f-measure of 92%. This measure is the harmonic mean of precision and recall, each of which assumes high values for that class. The recall is 1.0: this means that every work place was labeled correctly as belonging to this class. This is also confirmed by the precision value where 86% of the items labeled as *Professional&OtherPlaces* do indeed belong to that class, and a small percentage of other items were incorrectly labeled as work place. Similarly, 89% of *College&University* was labeled correctly, and only about 22% of other places are classified as educational locations. On the whole, the recall value for the other categories is on average around 70%, except for *Food* and *Nightlife&Spot* that presents misclassification for nearly half the time.

In Figure 9, for each category we show the error bars that correspond to standard deviations in the performance of the six

Category	P	R	FM
Art&Entertainment	0.64	0.68	0.66
College&University	0.78	0.89	0.83
Food	0.58	0.48	0.52
Nightlife&Spot	0.59	0.53	0.56
Outdoors&Recreation	0.67	0.69	0.68
Professional&OtherPlaces	**0.86**	**1.00**	**0.92**
Shop&Service	0.58	0.68	0.63
Travel&Transport	0.63	0.70	0.66

TABLE I
PRECISION (P), RECALL(R) AND F-MEASURE(FM) OF THE MOST PERFORMING CLASSIFICATION ALGORITHMS FOR EACH OF THE EIGHT LOCATION CATEGORIES.

classifiers over each class. These bars are short, meaning that the experimental measurements dispersion of all the classifiers on a specific category is small. This is particularly evident looking at the plot in Figure 10, where all the classifiers tend to infer well the same categories as *Professional&OtherPlaces* and *College&University*. The learners have an average accuracy of about 67%. The less performing is the K* with 64% of average accuracy.

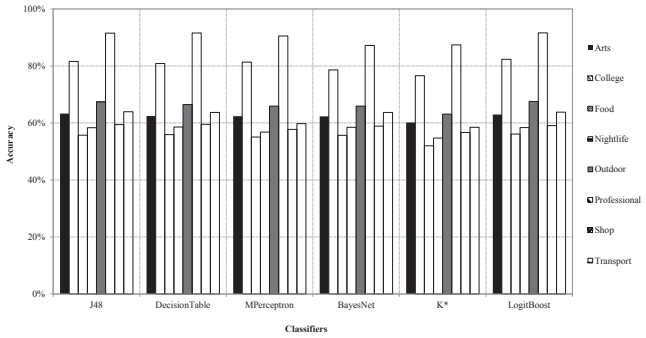


Fig. 10. Accuracy of the six classifiers over the eight location categories in solving the Binary Problem.

Classifiers are able to better discriminate the categories of places where users behavior is more stable and regular. Moreover, the accuracy rate is high for categories in which it is possible to identify a predominant user pattern.

Profiling problem. A more difficult problem is to choose the category of a place among a set of categories. The problem is formulated as a multinomial classification problem:

Given the set of category C*, which is the category* $c \in C$ *of a location?*

We consider three categories of places representative of the different human activities that typically characterize the everyday life of people. These categories are *Food&Drink*, *Leisure* and *Work*. In particular, the category *Food&Drink* includes all the places in which people eat and drink, like day-night bars and cafes (e.g., coffee shop, Whisky Bar, Wine Bar), restaurant (e.g., Italian, Chinese, Pizza), pub and so on. This category is the union of the Foursquare categories *Food* and *Nightlife&Spot*. A *Leisure* place is a place in which people spend time free from the demands of work or duty, where one can rest, do shopping, travel, enjoy hobbies or sports. This category groups the Foursquare categories *Arts&Entertainment*, *Outdoors&Recreation*, *Shop&Service* and *Travel&Transport*. Conversely, the category Work includes place in which people spent time in business, work, and education. It includes the Foursquare categories *College&University* and *Professional&OtherPlaces*.

The results of our analysis show that the aggregated sets present similarities between their spatial-temporal patterns. For instance, the visits at recreation places, art and entertainment locations present similar properties, attracting opportunistic visitors, generally in similar time period. On the contrary, the time that users spend in the work and educational places is different and longer. Moreover, public areas and transportation points, as well as food places are visited by a greater number of people, without periodic behavior, than the work places. As evidence of what is said, the following Figures 11 and 12 show two features: the daily time that users spend in a place (Daily User Stay), and the average number of visitors per place, of all the Foursquare categories.

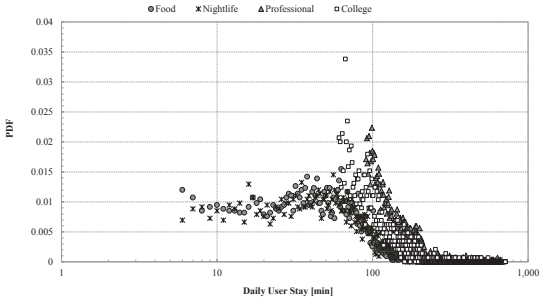


Fig. 11. Probability Distribution Function of the daily time that users spent in Food, Nightlife&Spot, College&University and Professional&OtherPlaces locations.

In Figure 11 we plot the PDF of Daily User Stay, of the pair of categories *Food* and *Nightlife&Spot*, grouped in *Food&Drink*, and the couple *College&University* and *Professional&OtherPlaces*, aggregated in *Work*. Notice that *Food* and *Nightlife&Spot* appear with similar temporal distributions, in which half of places have a mean daily time of stay of less than 60 minutes, and another 30% between 60 and 90 minutes. Conversely, *College&University* and *Professional&OtherPlaces* exhibit longer times to stay: in all this kind of locations the users spend at least 1 hour.

Figure 12 clearly shows that the places associated with the category *Food&Drink* (i.e., *Food* and *Nightlife&Spot*) are visited by an average of 7-8 users, while the Leisure places (i.e., *Arts&Entertainment*, *Outdoors&Recreation*, *Shop&Service* and *Travel&Transport*) have a number of users exceeds 15, and finally in the Work places (i.e., *College&University* and *Professional&OtherPlaces*) the number of visitors has averaged 5.

The input of the classifiers is an annotated training dataset, in which we have three groups of instances, each one belonging to one category set and labeled either with *Food&Drink*, *Leisure* or *Work*. In solving the profiling problem, we used the

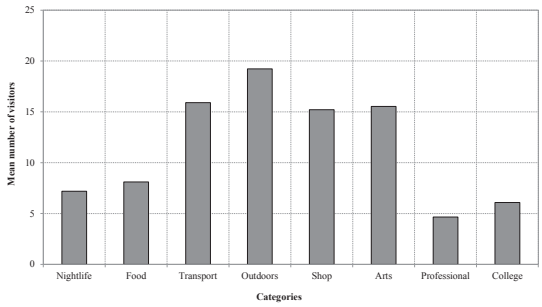


Fig. 12. Mean number of visitors for the eight location categories.

six mining algorithms, as for the binary problem. In Figure 13 it can be observed that the labels can be predicted with a mean accuracy of about 64%. The highest performing classifier is LogitBoost with an accuracy of 66%, while BayesNet and the K* discriminate instances with accuracy slightly exceeding the 60%.

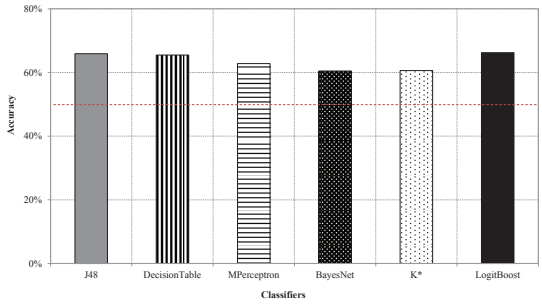


Fig. 13. Accuracy of the six classifiers in solving the Profiling problem.

Table II shows the values of precision (*P*), recall (*R*), and f-measure (*FM*) as they result from the LogitBoost classifier. The precision values indicate that for the categories *Food&Drink* and *Work* about 70% of items are correctly labeled as belonging to the correct class. *Leisure* presents a precision of 64%, and it has a good recall equals to 72%, overcome by *Work* that has the best recall value. In particular, 84% of work and educational places is identified as *Work*. On the contrary, the proportion of places belonging to *Food&Drink* that are correctly identified as such is not very high, this because they are confused with *Leisure* places. This is a consequence of the fact that the two classes present some similarities: for instance, both the categories of places attract visitors in similar time period, e.g., Casino belonging to *Leisure* and Nightclub belonging to *Food&Drink* are visited during night time.

Category	P	R	FM
Food&Drink	0.71	0.36	0.48
Leisure	0.64	0.72	0.68
Work	0.68	0.84	0.75

TABLE II
PRECISION (P), RECALL(R) AND F-MEASURE(FM) OF FOOD&DRINK, LEISURE AND WORK.

VII. CONCLUSION

In this paper we addressed the problem of automatic labeling the places of a city knowing approximate information coming from geo-tagged tweets. We formulated the problem as a supervised classification task. We introduced a novel approach to discovery spatial-temporal patterns from dynamics of human activity. Thanks to the extraction of this information we characterized each place with a set of machine learning features. Our results show that the proposed methodology allows to (i) infer if a place belongs to a certain category or not; and (ii) to choose the category of a place among a set of categories. Future work include the integration of such a framework in a recommender system as well as the study of different Twitter users profiles based on the patterns discovered with our technique.

ACKNOWLEDGMENT

The work presented in this paper has been partially supported by European Commission, European Social Fund (ESF), Regione Calabria, and COST program Action IC1305, 'Network for Sustainable Ultrascale Computing (NESUS)'. We acknowledge the support of the Engineering and Physical Sciences Research Council through grant GALE (EP/K019392).

REFERENCES

[1] E. Achtert, T. Bernecker, H.-P. Kriegel, E. Schubert, and A. Zimek. Elki in time: Elki 0.2 for the performance evaluation of distance measures for time series. SSTD, pages 436–440, 2009.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, 1999.

[3] D. Chalmers, S. Fleming, I. Wakeman, and D. Watson. Rhythms in twitter. SocialObjects, 2011.

[4] Z. Chen, Y. Chen, S. Wang, and Z. Zhao. A supervised learning based semantic location extraction method using mobile phone data. In *CSAE*, pages 548–551, 2012.

[5] T. Fujisaka, R. Lee, and K. Sumiya. Exploring urban characteristics using movement history of mass mobile microbloggers. HotMobile, pages 13–18, 2010.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[7] S. Kinsella, V. Murdock, and N. O'Hare. "i'm eating a sandwich in glasgow": Modeling locations with tweets. SMUC, pages 61–68, 2011.

[8] J. Krumm and D. Rouhana. Placer: Semantic place labels from diary data. UbiComp, pages 163–172, 2013.

[9] L. Liao, D. Fox, and H. Kautz. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.

[10] R. Montoliu, J. Blom, and D. Gatica-Perez. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, 62(1):179–207, 2013.

[11] S. Wakamiya, R. Lee, and K. Sumiya. Urban area characterization based on semantics of crowd activities in twitter. In *GeoSpatial Semantics*, pages 108–123. 2011.

[12] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *KDD*, pages 520–528, 2011.

[13] Y. Zhu, Y. Sun, and Y. Wang. Nokia mobile data challenge: Predicting semantic place and next place via mobile data. In *Mobile Data Challenge Workshop*, 2012.