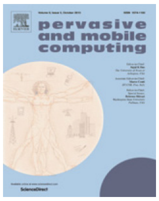
Pervasive and Mobile Computing 33 (2016) 91–107

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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



Mining human mobility patterns from social geo-tagged data

Carmela Comito[a,*], Deborah Falcone[b], Domenico Talia[b]

[a] *ICAR-CNR, Rende(CS), Italy*

[b] *DIMES, University of Calabria, Rende(CS), Italy*

ARTICLE INFO

Article history:
Received 13 November 2015
Received in revised form 24 March 2016
Accepted 8 June 2016
Available online 16 June 2016

Keywords:
Trajectory pattern mining
Geo-social data
Human mobility

ABSTRACT

Online social networks allow users to tag their posts with geographical coordinates collected through the GPS interface of smart phones. The time- and geo-coordinates associated with a sequence of posts/tweets manifest the spatial–temporal movements of people in real life. This paper aims to analyze such movements to discover people and community behavior. To this end, we defined and implemented a novel methodology to mine popular travel routes from geo-tagged posts. Our approach infers interesting locations and frequent travel sequences among these locations in a given geo-spatial region, as shown from the detailed analysis of the collected geo-tagged data.

1. Introduction

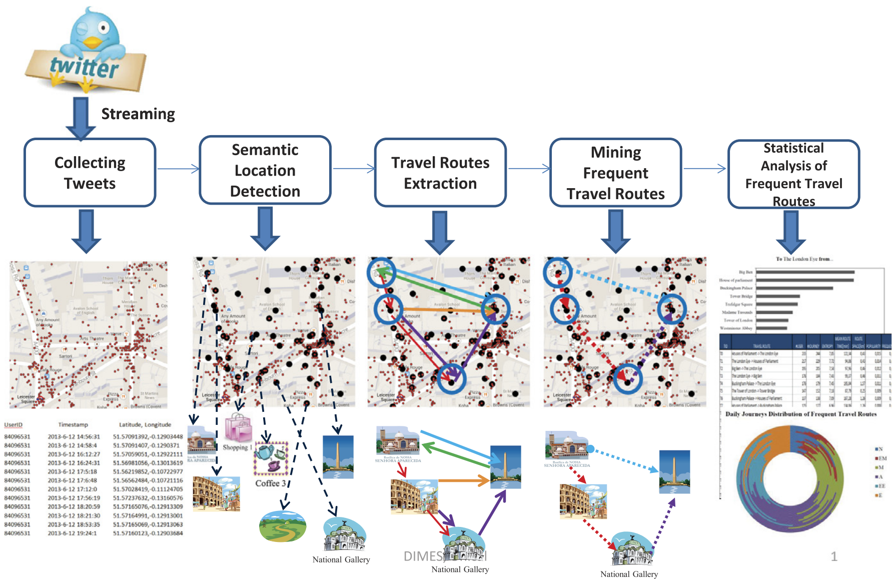
The ability to associate spatial context to social posts is a popular feature of the most used online social networks. For example, Facebook and Twitter exploit the GPS readings of users phones to tag posts, photos and videos with geographical coordinates. According to this view, social network users traveling and visiting a set of locations can produce a huge amount of geo-location data that embed extensive knowledge about human dynamics and mobility behaviors within urban context.

The potential to harness the rich information provided by geo-tagged social data may impact many areas including urban planning, intelligent traffic management, route recommendations, security and health monitoring. As such, a tremendous opportunity exists to develop effective tools to analyze and exploit those very large-scale spatial–temporal data. Moving toward this direction, the work presented in the paper aims to analyze the time- and geo-referenced information associated with online posts to detect typical trajectories and discover common behavior, i.e. patterns, rules and regularities in moving trajectories. The basic assumption is that people often tend to follow common routes: e.g., they go to work every day traveling the same roads. Thus, if we have enough data to model typical behaviors, such knowledge can be used to predict and manage future movements of people. In particular, the objective of this study is to provide the top interesting locations and frequent travel sequences among these locations, in a given geo-spatial region. For interesting locations we mean culturally important places, such as The National Gallery or Buckingham Palace in London (i.e., popular tourist destinations), and commonly frequented public areas, such as shopping malls/streets, restaurants and cinemas.

To effectively perform this kind of analysis, an integrated approach able to support the whole knowledge discovery process is needed. Although some approaches extracting popular itineraries from geo-tagged social data have been proposed, the existing literature lacks of a comprehensive system addressing all the issues involved in the process, from data acquisition to trajectory patterns characterization. To this aim we defined and implemented a novel methodology to mine popular travel routes from geo-tagged posts of a urban area such as those collected from Twitter. The proposed methodology

[*] Corresponding author.
E-mail addresses: comito@icar.cnr.it (C. Comito), dfalcone@dimes.unical.it (D. Falcone), talia@dimes.unical.it (D. Talia).

http://dx.doi.org/10.1016/j.pmcj.2016.06.005
1574-1192/© 2016 Elsevier B.V. All rights reserved.


twitter
Streaming
Collecting Tweets
Semantic Location Detection
Travel Routes Extraction
Mining Frequent Travel Routes
Statistical Analysis of Frequent Travel Routes
Shopping 1
Coffee 3
National Gallery
DIMES
National Gallery
National Gallery
To The London Eye from...
Daily Journeys Distribution of Frequent Travel Routes
1

UserID	Timestamp	Latitude, Longitude
84096531	2013-6-12 14:56:31	51.57091392,-0.12903448
84096531	2013-6-12 14:58:4	51.57091407,-0.1290371
84096531	2013-6-12 16:12:27	51.57059051,-0.12922111
84096531	2013-6-12 16:24:31	51.56981056,-0.13013619
84096531	2013-6-12 17:5:18	51.56219852,-0.10722977
84096531	2013-6-12 17:6:48	51.56562484,-0.10721116
84096531	2013-6-12 17:12:0	51.57028419,-0.11124705
84096531	2013-6-12 17:56:19	51.57237632,-0.13160576
84096531	2013-6-12 18:20:59	51.57165076,-0.12913309
84096531	2013-6-12 18:21:30	51.57164991,-0.12913001
84096531	2013-6-12 18:53:35	51.57165069,-0.12913063
84096531	2013-6-12 19:24:1	51.57160123,-0.12903684

Fig. 1. Methodology.

consists of various phases allowing to collect tweets, detect locations from them, identify travel routes between such locations, mine *frequent travel routes* using sequential pattern mining and extract spatial–temporal information for each of those routes to capture the factors that may drive users' movements. In particular, for all the frequent patterns, we compute a set of daily snapshots including the visited places, the movements among them, and the duration of the visits at each location.

The main contributions of the paper can be summarized as follows.

- We propose an effective approach for detecting relevant semantic locations from geo-tagged posts. The novelty of our work consists in identifying place semantics through a supervised approach purely based on spatial–temporal features such as stay duration, time of day, place popularity.
- We formulate the mobility pattern mining problem introducing a novel concept of travel route expressed as a sequence of consecutive visits to locations where tweets have been posted.
- We performed a single-to-crowd analysis, mining both user-centric behavior and common patterns of the city as a whole. We aggregate the individual trips to form a graph representing collective tourist behavior, and generate intra-city travel itineraries from the graph.

In the paper one scenario is introduced as a case study. This scenario focuses on a real-world dataset concerning tweets posted within an urban area of the city of London. As result of the analysis, useful insights in terms of common movements followed in the city are provided including the most popular itineraries followed by people's travels and the spatio-temporal distribution of such travels.

The rest of the paper is organized as follows. Section 2 presents the methodology. Section 3 describes the statistical analysis of frequent travel routes together with the experimental results performed on a real-world dataset which we exploit as case study. Section 4 overviews related works. Finally, Section 5 concludes the paper and proposes some further research issues.

2. A methodology to extract frequent travel routes from geo-tagged social data

Fig. 1 outlines the main steps of the methodology designed to identify people trajectories from geo-referenced posts of online social networks. The first step is the collection of geo-tagged data (e.g., from Twitter). The second step is semantic location detection. Locations are detected through dense clustering that allows to cluster GPS coordinates into specific places and associate them to Foursquare categories when available. After that, we extract trajectories and then mine frequent travel routes using sequential pattern mining. Finally, from all the frequent patterns, we extract spatial–temporal features so as to capture the factors that may drive users movements.

Here we describe in detail each step of the proposed methodology.

2.1. Collecting tweets

To collect tweets, we implemented 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 triple according to the following definition:

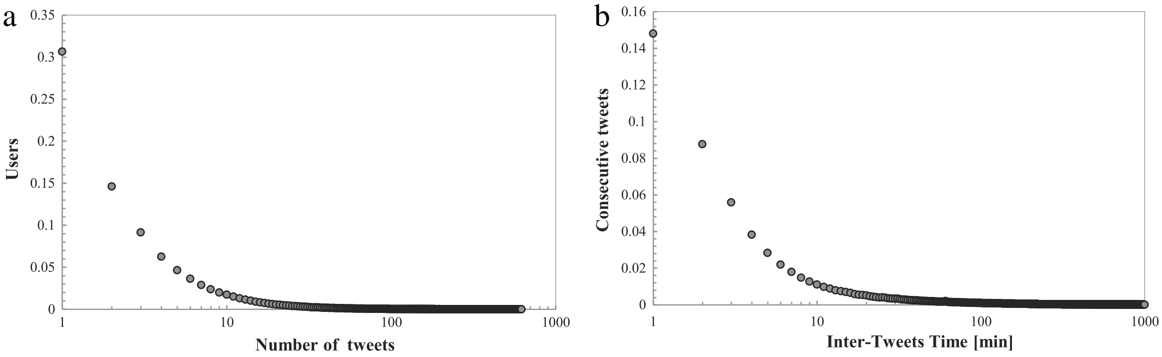


Fig. 2. Probability distribution function of number of monthly tweets per user (a) and of the time elapsed between consecutive tweets (b).

Definition 1 (*Geo-Tagged Tweet*)**.** A geo-tagged tweet $tw \in TW$ is characterized by the user u who tweeted, a location l from where tw has been posted and a timestamp, t, at which it has been posted. The location is identified by a pair of geographic coordinates $l = (x, y)$, latitude and longitude, respectively. Accordingly, a geo-tagged tweet can be defined as a triple $tw = (u, l, t)$.

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 months started in June 2013 and ended in November 2013. 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 the 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. Fig. 2(a) 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 could depend on the fact that many users are tourists who occasionally visit the city. 21% of users are more active making more than 10 tweets but less than 100, finally very few users, just 1%, post more than 100 tweets per month.

A similar pattern arises considering the time elapsed between successive tweets. Fig. 2(b) shows that about 40% of tweets are posted with high frequency (i.e., with an intertime of 10 min). However the other 60% of inter tweet time intervals have a length that varies in a very large range of values. On the other hand, only 28% of tweets are posted with a frequency greater than 3 h.

To exploit the temporal patterns for our classification task we divide the day into six different time slots, formally specified as follows:

Definition 2. TS is a finite set of timeslots with $|\text{TS}| = 6$. Each $ts \in \text{TS}$ is a time-object of varying time duration belonging to a day. TS = {N, EM, M, A, EE, E} where:

N = Night[12:00 am–05:59 am];
EM = EarlyMorning[06:00 am–09:59 am];
M = Morning[10:00 am–01:59 pm];
A = Afternoon[02:00–05:59 pm];
EE = EarlyEvening[06:00 pm–08:59 pm];
E = Evening[09:00 pm–11:59 pm].

On the basis of this definition, we specify a mapping function TS(tw) that associates the corresponding time slot to the timestamp of a tweet.

2.2. Semantic location detection

The *semantic location detection* step consists of two main phases: (i) location detection and (ii) semantic category association to the detected locations so as to obtain *semantic locations*.

2.2.1. Location detection

Due to imprecision of GPS, with a positional of average 10 m, a specific place in the city might be represented by slightly different GPS coordinates. To overcome this, we *cluster the locations* of the geo-tagged tweets so that each place is identified

[1] http://semiocast.com/.

Table 1
Performance of different location detection methods.

Method	Accuracy (%)	Detected locations (%)
Ad-hoc optics	**90.8**	**84.61**
DBSCAN	88.85	82.53
Mean-shift	87.27	81.66
lat/lon grids	82.23	76.13
Mapping to POI OSM	88.04	75.32
Mapping to POI 4SQ	82.04	70.53
Mapping to POI LP	83.04	56.81

by a pair of geographic coordinates. To this aim, we used a dense clustering algorithm. The result of the clustering is a model composed of a set of clusters; each of such cluster corresponds to a geographic region that actually is a *dense region* visited by many users.

Using a density-based clustering algorithm, we formulate a tree-based hierarchy by grouping together close geo-locations into some geo-spatial regions (set of clusters) in a divisive manner. To this aim we modified OPTICS [1], an algorithm for finding density-based clusters in spatial data. The output of the clustering algorithm has a hierarchical structure where a node on a level of the tree (that is a cluster of locations) represents a larger region for its descendant nodes and can be used to represent them obtaining this way a lower granularity in the location identification process. In general, regions covered by clusters on different levels of the hierarchy might stand for various semantic meanings, such as a city, a district and a community. In other words, a location might belong to multiple clusters based on the different region scales it falls in. To detect significant locations in a city, the algorithm parameters are set such as to aggregate the geo-tagged locations at the extend of venues in the city. To this aim we used a clustering neighborhood radius of about 50 m as it is reasonable to take this value for the spatial extension of typical locations in city environments.

We evaluated our location detection solution against the most representative approaches in literature. Specifically, for comparison purpose, we run some related techniques (listed below) over the reference dataset of tweets:

- DBSCAN, that is among the most widely used algorithms to identify locations from geo-tagged data (see Refs. [2–4]).
- Mean-shift clustering, which is the approach used in [5–8]. For comparison purposes, we run mean-shift over the reference dataset with the same parameter settings as in [5] since as in our case the extent of locations are typical venues in London city.
- The lat/lon Grids method proposed in [9]. The method divides the latitude and longitude space into grids of approximately 30 squared meters. Each grid corresponds to a location.
- Mapping to POI, in which tweet geo-coordinates are directly mapped to a set of pre-defined locations, the so-called POI. POI are obtained by online sources (e.g., Foursquare, Yahoo Maps, OpenStreetMap) (e.g., [10]) or city guide, e.g. Lonely Planet [11]. To the aim of the evaluation we implemented this approach by considering three POI sources, OpenStreetMap (OSM), Foursquare (4SQ) and Lonely Planet(LP). The limitation of this approach is that it excludes from the analysis other important locations that are not classified as POI in the reference source and from where many social data have been generated.

The comparison has been performed by introducing the following metrics: (i) the average accuracy expressed as percentage value, calculated using the Haversine distance between the estimated coordinates (e.g., cluster/grid centroids or mapped POI) and the real place coordinates (ground truth), as due to clustering/mapping errors locations may not be correctly identified; (ii) the percentage of detected locations, as due to clustering/mapping errors some locations may not be detected. As ground truth dataset we refer to Google Maps.

The main results of such experiments are summarized in Table 1. Results show that our location detection approach outperforms related works in terms of both accuracy and percentage of locations identified. This means that we detect the real locations from where tweets have been posted with higher accuracy compared to the related works, introducing on average a very low error in the identification of the unique geo-coordinates of the locations. Specifically, we achieved an average accuracy of about 91% versus 89% and 87% of DBSCAN and mean-shift, respectively. On average we detect about 85% of the overall locations. DBSCAN and mean-shift methods identify a quite similar list of locations and perform very similarly detecting on average about 82% of the overall locations. As expected, the approaches based on POI achieve the worst performance due to the approximation introduced in the mapping process, with the mapping to OpenStreetMap producing better results both in terms of accuracy and number of locations identified.

2.2.2. Category association

While some of the most recent location-based social networks such as Foursquare allow a user to explicitly indicate the place category he/she is at, in most of the other social networking tools, a location is simply represented by latitude–longitude coordinates. So that without user input the identification of the place the user is at is complex. However, the semantics knowledge of the category of a place (home, office, museum) is a significant information that can be added to the traditional latitude–longitude coordinates used by online social networks to represent a location: it could allow to infer users common interests, to improve activity prediction and ultimately mobile user recommendation and advertisement.

Table 2
Precision (P), Recall (R) and F-Measure (FM) of the LogitBoost for each of the eight location categories.

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

The automatic labeling of locations is the objective of this step of the methodology. Accordingly, once the locations (clusters) are detected we identify place semantics. 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. We formulate the problem as a supervised learning task where we want to associate the detected locations to the most-likely place category (e.g., Restaurant, School, Airport). Specifically, we assume that we can associate some locations to the most-likely category extracted from the Foursquare database of categories and coordinate associations. This will be the class attribute exploited by the supervised learning algorithm which is able to automatically label places. We used as ground truth labels a Foursquare database of London retrieved by the Foursquare API. This database contains 39,304 Foursquare venues. Each location l is represented by a couple: $l = \langle (lat, lon), 4sq_l \rangle$ where in addition to the pair of geographic coordinates is specified the category, $4sq_l$, of the most-likely Foursquare venue that can be associated to l. We refer to the eight top categories of Foursquare, that are Arts & Entertainment, College & University, Food, Nightlife & Spot, Outdoors & Recreation, Professional & OtherPlaces, Shop & Service, Travel & Transport.

As compared to a raw GPS point, each location detected with our methodology carries a particular semantic meaning, such as the shopping malls we accessed or the museums we visited, etc. Moreover, we made a more fine grain classification, retrieving by the Foursquare API, the precise name of places, e.g. London Eye, Buckingham Palace, National Gallery and so on. Details of our approach to infer semantic category of locations can be found in [12].

After a comparative analysis of classification algorithms, we opt for LogitBoost using the implementation included in Weka [13]. In order to estimate the accuracy of our prediction algorithm, 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).

Fig. 3 shows the accuracy achieved for each Foursquare category by the classifier and two baselines, linear regression and mapping to Foursquare venues. This last baseline is the approach presented in [10], where the most-likely Foursquare venue that can be associated to the geo-coordinates of tweets is retrieved by simply relying on geographic proximity. Results highlight the good accuracy achieved by the classifier that outperforms the baselines. In particular, the gain achieved by the classifier confirms that spatial–temporal patterns allow to categorize a place with a higher accuracy than methods based only on space distances, like the approach used in [10]. The mapping to Foursquare venues reaches results close to the ones obtained by the classifier, even if slightly worse, only for categories including a large number of Foursquare places (e.g. Food, Night, Arts). Furthermore it can be noted that, the accuracy of the classifier is significantly higher than that of the linear regression in all cases.

From Fig. 3 can also be noted that the highest 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. On the contrary, work places tend to be visited according to a unique temporal pattern. In general, the accuracy rate is high for categories in which it is possible to identify a predominant user pattern. Classifier is, therefore, able to better discriminate the categories of places where users behavior is more stable and regular.

In Table 2 we show the values of the validation metrics precision (P), recall (R), and f-measure (FM) of the classifier 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.

For the aim of this work we identified locations with a geographical extent of venues. Analyzing the dataset of twitter of London, with the semantic location identification phase we clustered 70 177 locations. As will be detailed in Section 3.1, after the clustering and the Foursquare mapping, London Eye, Buckingham Palace and Harrods are the top three popular locations and the top popular locations belong to the *OutdoorsRecreation* and *ArtsEntertainment* categories.

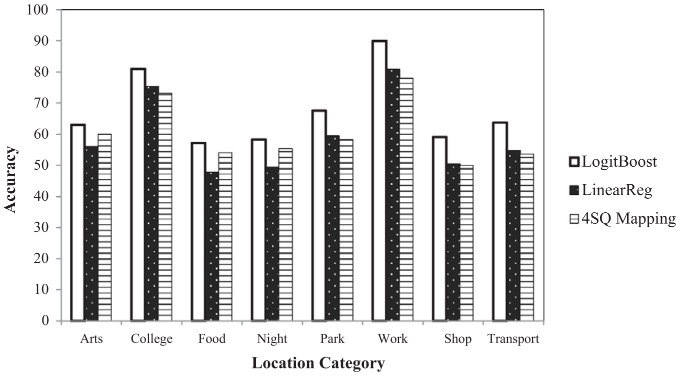


Fig. 3. Accuracy of the classifier, compared to two baselines: Linear Regression and Mapping to Foursquare venues.

2.3. Travel routes generation

This section discusses the extraction of trajectories, referred to as *travel routes*, from the collected geo-tagged tweets.

Given a set of geo-tagged tweets, we first identify the set of locations visited from Twitter users, and then, extract travel routes as temporally ordered sequences of places visited by users. Specifically, in a day d a user u might visit one or more locations in a city. For each user we compute his daily travel routes, formalized as follows:

Definition 3 (*Travel Route*)**.** A travel route is a spatio-temporal sequence of visited locations by a user u according to temporal order during the same day d.

$$\mathrm{TR}_{u,d} = v_{l_0,t_0} \longrightarrow v_{l_1,t_i} \longrightarrow \cdots \longrightarrow v_{l_n,t_n}$$

t_i denotes the timestamp of the first tweet posted in the location, such that $t_i < t_{i+1}$ (for each) $\vee_{0<i<N}$.

A visit to a location is, thus, the key concept to the daily travel route definition and is characterized by: (i) u, the user who visits the location l; (ii) $(tw_{\text{first}}, \ldots, tw_{\text{last}})$, a sequence of *consecutive tweets* that u posts in l before moving to another place; and (iii) $\Delta v = \mathrm{TS}(tw_{\text{last}}) - \mathrm{TS}(tw_{\text{first}})$, the duration of the visit that is equal to the difference between the timestamp of the last and the first tweet posted during the visit.

A central concept in the definition of a visit to a location is the one about consecutive tweets based on the time slots introduced above.

Definition 4 (*Consecutive Tweets*)**.** Two tweets tw_1 and tw_2, temporally ordered, are consecutive iff:

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

where $succ(\mathrm{TS}(tw_1))$ is the successive time slot of tw_1 and $\delta(tw_1, tw_2)$ 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 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). A daily trajectory, therefore, contains distinct visits to locations.

For each visited location l in a given travel route $\mathrm{TR}_{u,d}$ we compute the daily time spent in it by user u during the day d, indicated with $tIn_l^{u,d}$. This time is composed of: (i) the overall visits duration; and (ii) the overall time for real movements of u from l. Formally:

$$tIn_l^{u,d} = \sum_{v_l \in \mathrm{TR}_{u,d}} \Delta v_l + \sum_{v_{l'} \in \mathrm{TR}_{u,d}} (m_{l',l}). \tag{2}$$

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

Fig. 4(a) and (b) show two features: the daily time that users spend in a place (*Daily User Stay*), and the average number of users per place, of all the Foursquare categories. The results of our analysis show that the visits at recreation places, art and entertainment locations present similar properties, attracting opportunistic users, generally in similar time period. On the contrary, the time that users spend in 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. In particular, from Fig. 4(a) that plots the PDF of Daily User Stay of the location categories, one can note that *Food* and *Nightlife&Spot* appear with similar temporal distributions, in which half of places has a mean daily time of stay of less than 60 min, and another 30% between 60 and 90 min. Conversely, *College&University* and *Professional&OtherPlaces* exhibit longer

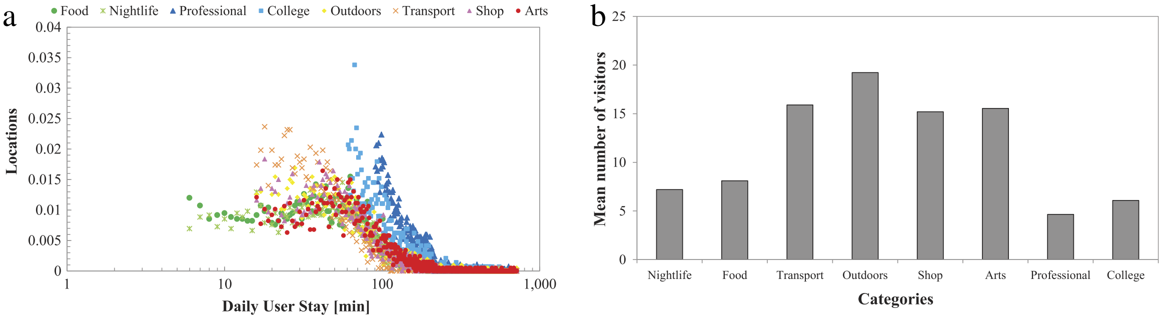


Fig. 4. PDF of the daily time spent in the location categories (a) and mean number of users for the location categories (b).

times: users spend in those places at least 1 h. Fig. 4(b) 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 users has averaged 5.

To the aim of this paper we extracted travel routes at the finest granularity detail, so travel routes among venues in the city. Precisely, we extracted about 455 422 travel routes of 292 195 users and involving 70 177 locations. The average cardinality of the extracted travel routes is 3.31.

2.4. Mining frequent travel routes

In this section we discuss the extraction of spatial–temporal patterns from the travel routes generated at the previous step.

Trajectory patterns can be used as a concise description of frequent behaviors, in terms of both space and time. Thus, the problem of trajectory pattern mining can be modeled as an extension of traditional sequential pattern mining and association rule mining. The idea is to determine sequences of places in the data which occur together frequently, and with similar transition times.

Given the travel routes generated at the previous step, the sequential pattern mining algorithm will find all the sequential patterns whose frequencies are no smaller than the minimum support. We refer to the mined frequent patterns as *trajectory patterns* or *frequent travel routes*.

Definition 5 (*Frequent Travel Route*). A trajectory pattern or a *Frequent Travel Route* is a sequence of locations frequently visited together, with a frequency no smaller than a minimum support *s*, that can be seen as a special association rule in the form:

$$\mathrm{FTR}_{u,d} = v_{l_0,t_0} \longrightarrow v_{l_1,t_i} \longrightarrow \cdots \longrightarrow v_{l_s,t_s}(s)$$

with $t_i < t_{i+1}$ (for each) $\vee_{0<i<s}$ and where *s* is the percentage of travel routes that contain it.

We adopt a two-phase approach for mining popular travel routes: (i) the first phase consists of applying sequential pattern mining on the location sequences; (ii) the second one consists of extracting the maximal frequent sub-sequences from all the frequent sequences mined. This second step is necessary in order to ensure that trajectories with large segments in common are not reported simultaneously. To this aim we propose an algorithm that extends the well-known PrefixSpan [14] algorithm to obtain only maximal frequent patterns.

In the following is described a toy example of frequent travel routes mining.

Example 1. Let us suppose to extract from the tweets of London the following travel routes:

TR_1 : *MillenniumBridge* → *StPaul'sCathedral* → *BigBen* → *TheLondonEye* → *TowerBridge* → *TateModern*
TR_2 : *BuckinghamPalace* → *MadameTussauds* → *BigBen* → *TheLondonEye* → *TowerBridge* → *WestminsterBridge* → *ParliamentSquare*
TR_3 : *BigBen* → *TheLondonEye* → *TrafalgarSquare* → *TempleofMithras*
TR_4 : *BigBen* → *TheLondonEye* → *TowerBridge* → *WestminsterBridge* → *ParliamentSquare*
TR_5 : *BuckinghamPalace* → *MadameTussauds* → *M&M'sWorld* → *TowerBridge* → *TateModern*.

Given the above 5 travel routes the maximal frequent pattern mining algorithm determines the following frequent travel routes with a minimum support fixed at 2:

FTR_1 : *BigBen* → *TheLondonEye* (frequency 4)
FTR_2 : *TheLondonEye* → *TowerBridge* (frequency 3)

FTR_3 : *TowerBridge* → *TateModern* (frequency 2)
FTR_4 : *WestminsterBridge* → *ParliamentSquare* (frequency 2)
FTR_5 : *BuckinghamPalace* → *MadameTussauds* (frequency 2).

Concerning our real evaluation of the London tweets, we mined 923 maximal frequent patterns to which we refer to as frequent travel routes. Among this patterns the most frequent ones are *BigBen* → *LondonEyes* with a frequency of 205 and *HouseofParliament* → *LondonEyes* with a frequency of 203.

2.5. Spatial–temporal features to characterize frequent travel routes

This section describes how to obtain daily snapshots and statistics on the mined travel routes. We define a set of features that exploit different information dimensions about users' movements including traveled paths, visited locations, the movements among them and the duration of the visits at each location. The most relevant features are listed below.

We denote the set of frequent travel routes in the Twitter dataset as $\mathcal{FTR}$, the set of popular locations $\mathcal{L}$, the set of category $\mathcal{C}$ and the set of users $\mathcal{U}$.

Number of journeys. The number of journeys along a travel route is the overall number of times that the route is traveled; it corresponds to the support of the frequent travel route *ftr* and it can be defined as follows:

$$\mathrm{Journeys}(ftr) = |\{ftr_i \in \mathcal{FTR} : ftr_i = ftr\}|. \tag{3}$$

Number of distinct travelers. The number of people who travels a route is indicative of its *popularity*. The number of travelers of a frequent travel route *ftr* can be expressed as follows:

$$\mathrm{Travelers}(ftr) = |\{u \in \mathcal{U} : \mathcal{J}_{ftr,u} \neq \emptyset\}| \tag{4}$$

where $\mathcal{J}_{ftr,u}$ is the set of journeys in the travel route *ftr* by user *u*.

Number of journeys per traveler. This feature describes the periodicity of users behavior along the trajectory. The feature is formalized as follows:

$$\mathrm{JourneysPerUser}(ftr) = \frac{\mathrm{Journeys}(ftr)}{\mathrm{Travelers}(ftr)}. \tag{5}$$

Travel route categories. This feature characterizes the travel route in terms of the categories of the locations crossed in the route. Formally can be expressed as follows:

$$\mathcal{C}(ftr) = \{c : \forall l \in \mathcal{L}_{ftr}, \mathrm{Category}(l) = c\} \tag{6}$$

where $\mathcal{L}_{ftr}$ is the set of locations in the travel route *ftr* and *Category*(l) represents the category of the location *l*.

Number of tweets. This feature accounts the total number of tweets that have been posted from the locations in a travel route:

$$TW(ftr) = |\{(u, l, t)\} \in \mathcal{TW},\ \forall l \in \mathcal{L}_{ftr}| \tag{7}$$

where $\mathcal{L}_{ftr}$ is the set of locations in the travel route *ftr*.

Entropy of a travel route. Entropy tells us how a route is visited, describing the distribution of the movements across the users: whether users tend to travel regularly the route at usual times or they transit in it without any regularity. For this purpose we use the Shannon Entropy:

$$H(X_{ftr}) = -\sum_{u=1}^{n} p(x_u) \log p(x_u), \quad \text{where } p(x_u) = f(ftr, u) \tag{8}$$

$f(ftr, u)$ is the user's proportion of journeys at the travel route *tr* and is defined as $f(ftr, u) = \frac{|\mathcal{J}_{ftr,u}|}{|\mathcal{J}_{ftr}|}$ where $\mathcal{J}_{ftr}$ is the whole set of journeys in the travel route *ftr* by user *u*, while $\mathcal{J}_{ftr}$ is the whole set of journeys in the travel route. We expect a small entropy for residential routes involving Professional 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 do journeys along the route *ftr*, but they have very few journeys. This user behavior is typical along touristic routes involving Arts&Entertainment or Nightlife&Spot places.

3. Evaluation study: understanding people mobility behavior

The aim of this section is to characterize and understand patterns and regularities in order to mine human mobility dynamics within an urban area.

In particular, we focus on the statistical characterization of the 923 frequent travel routes obtained after applying the mining methodology on the London tweets dataset. To this aim we exploit the features defined in the previous section.

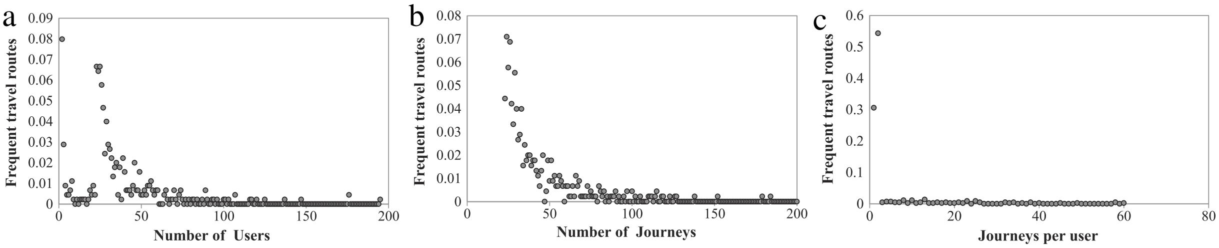


Fig. 5. Probability distribution function of number of distinct users (a) number of journeys (b) per frequent travel route and (c) number of journeys per user.

In the analysis we first investigated the characterization of aggregate human mobility and activity patterns (group behavior, e.g. tourists). Next, we focused on the characterization of individual mobility patterns, e.g., social interactions and travel patterns of people living in their constituent neighborhoods (e.g., residents). Accordingly, we distinguish among routes involving a large community of people and routes taken by a single person or a small group of people (individual mobility). The *collective* routes give us indication about crowd mobility and, thus, how people behave in the city of London, whereas *individual* routes characterize a given person, highlighting her/his daily mobility patterns and giving insights about her/his daily routines.

3.1. *Collective travel routes*

In this section we focus on 451 collective frequent travel routes. The overall analysis of traffic flows across time and space reveals that the majority of the frequent travel routes correspond to tourist movement patterns, as the visited locations are London's most important historical and cultural sites. Moreover, the spatial–temporal information featuring the travel routes, like the number of users traveling the route (popularity), the number of journeys (frequency) along the route, and the regularity of users behavior (entropy), outline the profile of touristic trajectories.

The average length of the frequent travel routes (the ones obtained after the sequential pattern mining process) is 2 whereas the average size of the overall travel routes is 3.2. This difference can be explained according to the following observation. 90% of frequent travel routes are followed by tourists who visit an average of 2.4 locations per day. This number is also confirmed in the literature as reported in the study [5] where the average length of touristic travel routes is 2, and in the study [3] where the authors explicitly say that it is not necessary to find out trajectories with long length as people would not visit many places in a trip and for this reason in that paper they work with 2-length trajectories.

The frequent travel routes obtained after the application of our methodology are frequent for two different reasons: (i) there are few users traveling the routes a high number of times and such routes highly characterize individuals expressing daily routines and activities; (ii) the routes are traveled by a high number of distinct persons; in this last case the routes are popular and could give indication about crowd mobility behavior in the city of London. Accordingly, the number of users traveling a specific route is indicative of its *popularity*. Fig. 5(a) shows the probability distribution function (PDF) of the number of distinct users per frequent travel route, highlighting that about 86% of travel routes is traveled by a number of distinct users greater than 10, with a peak when the number of users is 23. Among them, about 10% of trajectories are visited by more than 60 users. Fig. 5(b) shows the PDF of the number of journeys per travel route. This value gives information about the *frequency* of travel routes. The graph highlights a trend similar to the previous distribution. In fact, more than 80% of the travel routes are traveled a number of times ranging from 10 to 60, while there is a few number of travel routes (about 16%) with a very large number of journeys. Those trajectories cross the most popular tourist attractions. This is also evident for the result in Fig. 5(c). In fact, even if the volume of journeys and users per travel route is quite high, it is interesting to observe that the number of journeys per user is 1 or 2 for 85% of travel routes, as shown in Fig. 5(c). This depends on the fact that those travel routes are tourist and, thus, users occasionally travel across them, as shown in Table 3.

In Fig. 6(a) we plot the mean visit time duration of frequent travel routes, named *mean route time*. The graph shows that most of the visits (88%) lasts in a time period longer than half an hour and shorter than four hours and a half, with a peak around two hours. This result could be explained as follows. As the majority of journeys are relative to tourist routes, it is reasonable and quite realistic to assume that visit duration time is in between half an hour and 4 h for tourist trips: as an example, a visit to Big Ben on average could last around half an hour, whereas a visit to the National gallery could last 4 h. In general, tourist attractions have a visit duration time that could be in this range.

Fig. 6(b) shows that in 97% of travel routes the sequences of visited locations are within the small radius of 4 km. In particular, 64% of trajectories have a *route space*, i.e. the distance in km between the first and the last location in the path, of about 1 km. Only 2% of travel routes cover a distance greater than 4 km, reaching a maximum of 10 km.

Table 3 lists the 20 top popular travel routes, frequent travel routes ranked according to the popularity (the number of distinct visitors). As expected, most of the popular routes concern visits to world famous historic sights, including Houses of Parliament, Big Ben, Buckingham Palace, and other London attractions like the London Eye and Harrods. Accordingly, we refer to the popular travel routes in Table 3 as *tourist travel routes*, characterized by a high number of distinct persons

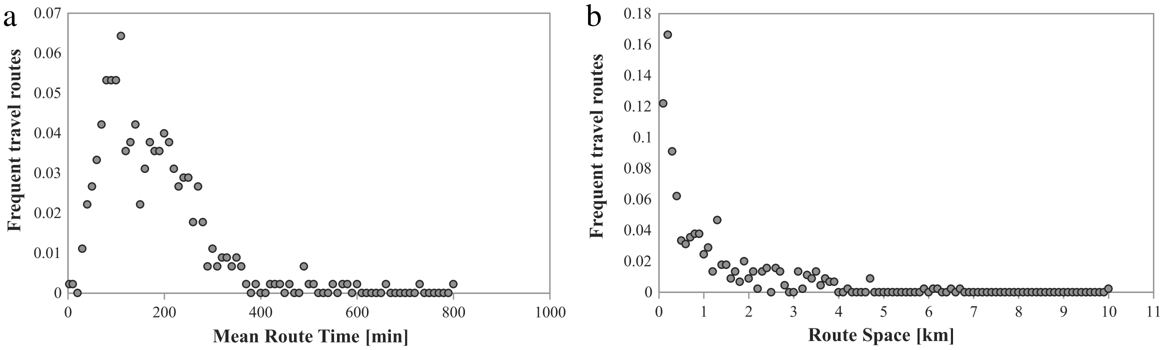


Fig. 6. Probability distribution function of mean route time (a) and route distance (b) per frequent travel route.

Table 3
The top 20 popular travel routes.

ID	Popular travel route	Tweets	Users	Journeys	Journeys/User	Entropy	Visit time (min)	Space (km)
T_0	Houses of Parliament ⇆ The London Eye	1121	452	473	1.05	7.78	108.2	0.43
T_1	Big Ben ⇆ The London Eye	924	371	389	1.05	7.51	98.66	0.46
T_2	Buckingham Palace ⇆ The London Eye	710	283	289	1.02	7.09	188.25	1.57
T_3	Tower of London ⇆ Tower Bridge	711	264	275	1.04	7.01	82.45	0.25
T_4	Buckingham Palace ⇆ Houses of Parliament	620	262	265	1.01	7.03	160.67	1.26
T_5	The London Eye ⇆ Tower Bridge	501	204	207	1.01	6.66	189.26	3.08
T_6	Buckingham Palace ⇆ Trafalgar Square	471	189	194	1.02	6.53	146.61	1.22
T_7	Big Ben ⇆ Buckingham Palace	441	189	190	1.01	6.56	135.2	1.21
T_8	Trafalgar Square ⇆ The London Eye	471	187	192	1.03	6.53	160.55	0.77
T_9	Trafalgar Square ⇆ National Gallery	423	181	186	1.03	6.42	66.12	0.08
T_{10}	St James's Park ⇆ Buckingham Palace	422	164	170	1.04	6.33	64.5	0.49
T_{11}	Houses of Parliament ⇆ Westminster Abbey	374	162	162	1	6.33	91.98	0.35
T_{12}	Houses of Parliament ⇆ Trafalgar Square	376	158	164	1.04	6.28	133.97	0.87
T_{13}	Buckingham Palace ⇆ Westminster Abbey	378	156	158	1.01	6.27	101.75	0.96
T_{14}	Madame Tussauds ⇆ The London Eye	456	152	153	1.01	6.24	276.48	3.28
T_{15}	Tower of London ⇆ The London Eye	346	132	137	1.04	6.02	193.66	3.05
T_{16}	Buckingham Palace ⇆ Tower Bridge	304	131	132	1.01	6.02	218.4	4.64
T_{17}	Westminster Abbey ⇆ The London Eye	293	126	128	1.01	5.96	118.62	0.76
T_{18}	Green Park ⇆ Buckingham Palace	285	121	123	1.02	5.9	55.51	0.29
T_{19}	Oxford Street ⇆ Piccadilly Circus	249	104	117	1.11	5.56	225.53	0.74

traveling along them (popular routes) and a high number of journeys along them. Important information is given by the high entropy values indicating that the behavior of visitors is not regular on the routes; in fact, according to our definition of the entropy feature, a high value means that there is a high number of distinct persons that travels the routes on average only once, as confirmed by values in column *Journeys/Users*. We observe that the district of Westminster presents the highest tourist density. The table also contains the average visit duration time and the route spatial distance (route space). We note that, on average, the travel routes last about 2 h and a half and cover a mean radius of 1.1 km, in accordance to a tourist profile and to Fig. 6(a) and (b). All these travel routes, could be used to plan an itinerary for a trip to London, providing information on attractions that are close/within walking distance of each other; grouping the most popular attractions by area to maximize tourists time and to avoid them to criss-crossing all over the city; and advertising the public transportation system and walking routes.

Fig. 7 shows the journey's temporal distribution of the top 20 popular travel routes, respectively during the hours of the day (a) and during the days of the week (b). Fig. 7(a) shows that, according to the tourist profile, the travel routes are traveled for a short time period during night (N) and early morning (EM), while they are traveled mostly during morning (M), afternoon (A) and early evening (EE). This is justified by the fact that generally the tourist attractions of a city are visited during daylight hours. Moreover, many of those attractions (e.g., museums, historical buildings, art galleries) have daily admission and opening times.

Fig. 7(b) presents a similar distribution of the number of journeys during the week for all the travel routes. The plot highlights that the routes have a number of journeys distributed throughout the week with peaks during week-ends.

Table 4 shows the top 20 popular locations, As expected, it contains the London's most famous tourist attractions. These locations are characterized by a high number of distinct visitors and by a high number of visits. The column *VisitPerDay* shows that the first 10 locations are visited daily a number of times greater than or equal to 10. The *FrequencyFTR* column specifies the travel routes to which the location belongs to, showing that the most popular locations present high frequencyFTR. For instance, The London Eye and Buckingham Palace are in 52 distinct travel sequences.

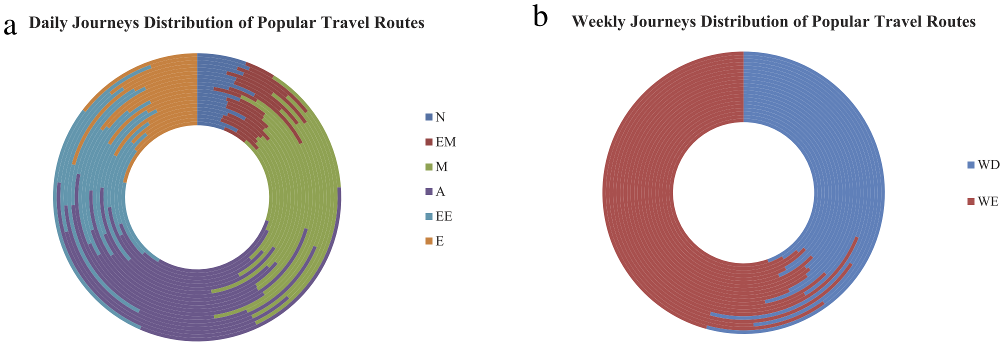


Fig. 7. Journey distribution of popular travel routes during: the hours of a day (a); the days of a week (b).

Table 4
The top 20 popular locations.

Popular location	Users	Visits	VisitPerDay	FrequencyFTR	Reachability	Incoming
The London Eye	1594	3323	21.58	52	24	25
Buckingham Palace	1284	2784	18.20	52	29	22
Westminster Abbey	996	1898	12.82	45	23	22
Houses of Parliament	922	1728	13.09	27	14	12
Trafalgar Square	890	1879	12.28	39	20	19
Tower Bridge	765	1419	9.72	27	12	15
Big Ben	737	1431	15.39	25	11	13
Hyde Park	650	1089	7.83	25	13	12
The Tower of London	539	988	7.06	22	12	10
British Museum	501	894	11.92	21	9	12
Oxford Street	458	892	9.91	22	12	10
Harrods	438	815	5.82	16	9	7
St Paul's Cathedral	374	655	4.85	17	10	7
Tate Modern	302	433	3.52	11	6	5
Madame Tussauds	278	368	3.29	9	5	4
Natural History Museum	235	309	2.45	8	5	3
St James's Park	219	326	2.91	7	4	3
National Gallery	216	315	2.92	6	2	4
Victoria and Albert Museum (V&A)	206	257	2.40	7	4	3
Piccadilly Circus	118	126	1.70	5	3	2

For each location, the table shows the number of other locations that can be reached (*Reachability*), and the number of places from where the specific location can be reached (*Incoming*). In agreement with *FrequencyFTR*, The London Eye and Buckingham Palace are the places that allow to reach/and are reached by the highest number of places. In Fig. 8 we present the reachability graph of the 10 top popular locations. The locations are represented as vertices connected by edges. An edge is a travel route between two locations, it has a direction and a weight that is the number of journeys along the specific direction. According to the values of Table 4, The London Eye and Buckingham Palace are the locations involved in the highest number of travel routes, thus they appear in the graph with the highest number of incoming and outgoing edges.

The set of extracted trajectory patterns represent a basic building block around which several urban computing applications can be implemented. Specifically, the reachability graph of Fig. 8 could be the core of a recommendation system that suggests routes and locations based on users' historical information. The key applications can be as follows.

- *Travel recommendations*. The mined top interesting locations and travel sequences can be exploited to recommend the best routes and itineraries that people can follow to visit a given location. Examples of travel recommendations include:
  1. Suggesting the shortest path between two locations: this information could be useful to users who want to visit the smallest set of places between origin and destination locations.
  2. Suggesting the most traveled path between two locations: this could be useful to users (e.g. tourists) who starting from the origin location want to visit other popular locations along the destination direction.
  3. Suggesting the best *k*-hop path: this could be useful to users (e.g. tourists) that starting from an origin location wish to visit the most traveled route composed of *k* locations.
- *Next location prediction*. Predict a list of places in which a user could move next, based on its recent movements and trajectory pattern models. This information could be useful to anticipate or prefetch possible services in that location.
- *Intelligent traffic management*. Predict traffic congestion patterns and improve the transportation model of a city, to reduce the wasted time due to traffic.
- *Movement-similarity analysis*. Estimate the similarity between users in terms of location histories so as to promote services for car sharing, car pooling, etc.

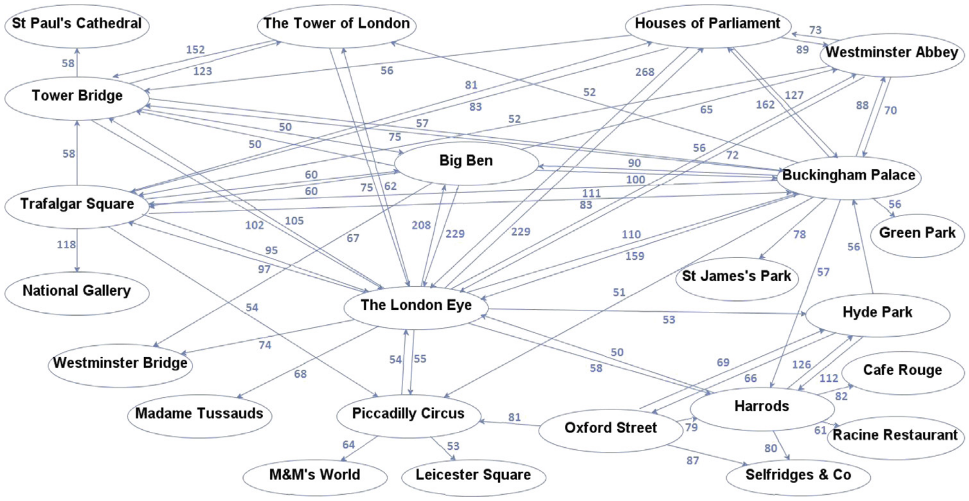


Fig. 8. Reachability graph.

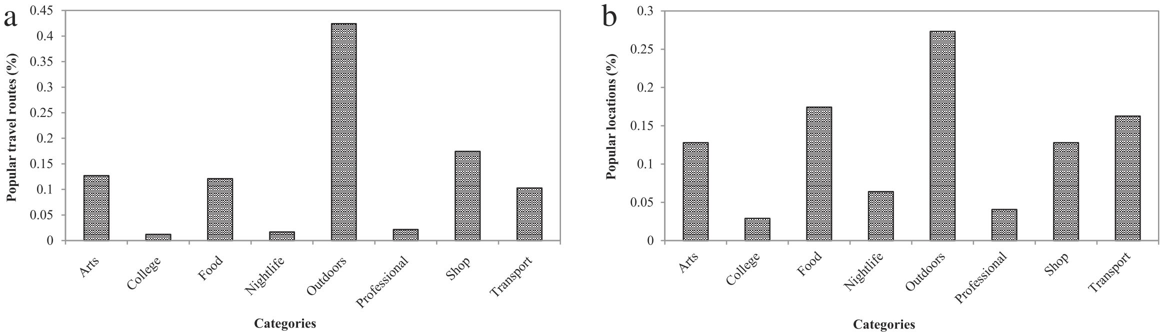


Fig. 9. Categories distribution of the popular travel routes (a) and of popular locations (b).

We determine place semantics according to Foursquare classification. Fig. 9(a) and (b) show the categories distribution to which belong the popular travel routes and the visited locations, respectively. Both the distributions highlight that the prominent activities of users are spending leisure time in the outdoor (e.g., park, square) or visiting landmarks (e.g. historic buildings, monuments). In fact 42% of routes have travel sequences among *Outdoor&Recreation* locations (Fig. 9(a)). Those locations represent 27% of the overall visited places (Fig. 9(b)). Over half of the travel routes (53%) visit places belonging to the categories *Food*, *Shop&Service*, *Arts&Entertainment*, and *Travel&Transport*. Only 5% of routes visit work, educational and night-life locations. A similar pattern arises considering the categories distribution of visited locations in Fig. 9(b). This evidence reinforces our hypothesis that most of the travel routes are tourist movements.

3.1.1. *Accuracy of the popular travel routes*

In this section we evaluate the accuracy of the extracted popular locations and travel routes assessing whether they are really popular. To reach this goal, as there are no similar benchmark to compare with, we test the extracted patterns with existing statistics. In particular, we refer to the rankings of different travel websites (e.g. TripAdvisor, Tripomatic and Viator), search engines (e.g. Google and Yahoo!Travel) and national tourist boards (e.g. VisitEngland).

We evaluated the accuracy of the extracted 20 top popular locations, comparing them against 10 lists of the most popular London attractions as ranked on 10 different web sources.[2] Attractions popularity in some of those sources is estimated based on the feedback scores provided by users (e.g., TripAdvisor, Google) others are official city tourist agency (e.g., City of London Information Centre) or private company (e.g., Tourist information UK, VisitLondon.com). The different nature of such web sources allows to assess whether people with different background tend to agree on the most popular attractions to some extent.

For each location *l* we compute the accuracy, Acc@*X*, using two different lists including 10 and 20 locations, respectively. We successfully rank a location *l* if it is in the top-X positions of the comparison list. Acc@*X* measures the fraction of times that

[2] www.google.com, www.tripadvisor.com, www.visitacity.com, www.tripomatic.com, www.visitengland.com, www.visitlondon.com, www.tourist-information-uk.com, www.britishtours.com,, www.planetware.com.

Table 5
Accuracy of the top 20 popular locations using (a) a list size of 10 (Acc@10) and (b) a list size of 20 (Acc@20).

(a)		(b)	
Popular location	Acc@10	Popular location	Acc@20
The London Eye	0.80	The London Eye	1.00
Buckingham Palace	0.90	Buckingham Palace	1.00
Westminster Abbey	0.70	Westminster Abbey	0.80
Houses of Parliament	0.80	Houses of Parliament	1.00
Trafalgar Square	0.60	Trafalgar Square	1.00
Tower Bridge	0.70	Tower Bridge	1.00
Big Ben	0.80	Big Ben	1.00
Hyde Park	0.20	Hyde Park	0.80
The Tower of London	0.90	The Tower of London	1.00
British Museum	0.60	British Museum	0.80
		Oxford Street	0.40
		Harrods	0.40
		St Paul's Cathedral	0.80
		Tate Modern	0.40
		Madame Tussauds	0.20
		Natural History Museum	0.60
		St James's Park	0.40
		National Gallery	0.80
		Victoria and Albert Museum (V&A)	0.60
		Piccadilly Circus	0.60

Table 6
Accuracy of the top 20 popular travel routes.

ID	Popular travel route	Accuracy
T_0	Houses of Parliament ⇆ The London Eye	1.00
T_1	Big Ben ⇆ The London Eye	1.00
T_2	Buckingham Palace ⇆ The London Eye	0.85
T_3	Tower of London ⇆ Tower Bridge	0.69
T_4	Buckingham Palace ⇆ Houses of Parliament	0.92
T_5	The London Eye ⇆ Tower Bridge	0.69
T_6	Buckingham Palace ⇆ Trafalgar Square	0.69
T_7	Big Ben ⇆ Buckingham Palace	0.92
T_8	Trafalgar Square ⇆ The London Eye	0.69
T_9	Trafalgar Square ⇆ National Gallery	0.69
T_{10}	St James's Park ⇆ Buckingham Palace	0.69
T_{11}	Houses of Parliament ⇆ Westminster Abbey	0.92
T_{12}	Houses of Parliament ⇆ Trafalgar Square	0.69
T_{13}	Buckingham Palace ⇆ Westminster Abbey	0.85
T_{14}	Madame Tussauds ⇆ The London Eye	0.08
T_{15}	Tower of London ⇆ The London Eye	0.69
T_{16}	Buckingham Palace ⇆ Tower Bridge	0.62
T_{17}	Westminster Abbey ⇆ The London Eye	0.92
T_{18}	Green Park ⇆ Buckingham Palace	0.23
T_{19}	Oxford Street ⇆ Piccadilly Circus	0.15

l is at the top-X positions in the other rankings. Table 5(a) and (b) confirm that London's most famous tourist attractions are highly ranked in both our lists of 10 and 20 top locations. As expected, the accuracy of the locations improves by considering lists of increasing size. For example, *Buckingham Palace* and *The Tower of London* are in the top-10 positions with 0.9 of accuracy, reaching 1 when we consider the list of 20 locations. Similarly, *The London Eye*, *Houses of Parliament* and *Big Ben* show very high accuracy in both lists, with Acc@10 $= 0.8$ and Acc@20 $= 1$.

Hyde Park is in the top-10 positions only in the 20% of the web sources, but its accuracy increases to 80% considering the top-20 locations. Shopping areas like *Oxford Street* and *Harrods* are not considered highly popular with an accuracy of 0.4. In summary, the tables show that historical buildings, important squares and museums are considered the best attractions in London, overcoming parks and shopping areas.

Likewise, we evaluated the accuracy of the extracted top travel routes. To the purpose, we compared our top popular travel routes against the ones listed in 13 travel itineraries recommended on the web[3] Each travel itinerary contains a different number of travel routes. For each travel route T_i we compute the accuracy as the fraction of times that T_i is included in travel itineraries. Table 6 reports the accuracy of the top 20 popular travel routes extracted by our methodology. Table shows that on average the accuracy is around 0.70: travel routes T_0 and T_1 are in all the itineraries (accuracy 1.0), followed

[3] www.tripadvisor.com, www.viator.com, www.visitacity.com, www.britishtours.com, www.londontourist.org, www.walklondon.com, www.freetoursbyfoot.com.

Table 7
The top 20 frequent individual travel routes.

ID	Frequent travel route	Tweets	Users	Journeys	Journeys/user	Entropy	Visit time (min)	Space (km)
T_0	BBar $\leftrightarrows$ $Home_0$	2401	1	250	250	0.04	588.99	14.71
T_1	Clapham Junction $\leftrightarrows$ ASDA Clapham Junction	1040	2	234	117	0.07	102.58	0.24
T_2	Red Lion $\leftrightarrows$ The Courtyard Theater	693	2	218	109	0.07	175.81	0.22
T_3	Northcliffe House $\leftrightarrows$ Logica Offices – Kings Place	610	1	217	217	0.05	270.81	6.04
T_4	Prince Arthur Pub $\leftrightarrows$ The Courtyard Theater	547	1	182	182	0.06	210.84	0.14
T_5	Cyclopedia → Wandsworth High Street	610	2	170	85	0.09	118.73	0.11
T_6	Prince Arthur Pub $\leftrightarrows$ Red Lion	440	1	160	160	0.09	222.44	0.35
T_7	Wickes Building Supplies $\leftrightarrows$ $Home_1$	3594	1	153	153	0.06	233.64	0.52
T_8	Marquess Tavern → $Home_2$	976	1	143	143	0.05	151.49	0.20
T_9	BBC Television Centre → Logica Offices – Kings Place	359	1	131	131	0.05	272.57	7.70
T_{10}	Westminster Wardens $\leftrightarrows$ The Talent Business	1780	1	117	117	0.07	604.63	3.47
T_{11}	06 St. Chad's Place → Burger King	483	3	114	38	0.14	93.84	0.29
T_{12}	$Home_0$ → BBar → $Home_0$	1506	1	100	100	0.04	1106.41	29.43
T_{13}	$Home_0$ → East Ham Underground Station → $Home_0$	801	1	81	81	0.04	1120.56	3.39
T_{14}	Hammersmith Tattoo → El Toro Restaurant	993	1	80	80	0.04	143.81	0.07
T_{15}	Clapham Junction → Bar Room Bar	373	2	76	38	0.10	108.04	0.92
T_{16}	$Home_0$ → East Ham Underground Station → BBar	753	1	71	71	0.04	683.97	15.95
T_{17}	$Home_0$ → East Ham Underground Station → BBar → $Home_0$	975	1	67	67	0.05	1132.83	30.67
T_{18}	The Crown → The Hemingford Arms	691	3	63	63	0.26	251.71	0.67
T_{19}	$Home_0$ → East Ham Leisure Centre → East Ham Underground Station → BBar → $Home_0$	773	1	52	52	0.05	1147.14	30.69

Table 8
Frequent travel routes of user *u*.

ID	Frequent travel route	Tweets	Journeys	Entropy	Visit time (min)	Space (km)
T_0^u	BBar → $Home_0$	1117	127	0.04	543.61	14.71
T_1^u	$Home_0$ → BBar	1284	123	0.04	634.37	14.71
T_2^u	$Home_0$ → BBar → $Home_0$	1506	100	0.04	1106.41	29.43
T_3^u	$Home_0$ → East Ham Underground Station → $Home_0$	801	81	0.04	1120.56	3.39
T_4^u	$Home_0$ → East Ham Underground Station → BBar	753	71	0.04	683.97	15.95
T_5^u	$Home_0$ → East Ham Underground Station → BBar → $Home_0$	975	67	0.05	1132.83	30.67
T_6^u	$Home_0$ → East Ham Leisure Centre → East Ham Underground Station → BBar → $Home_0$	773	52	0.05	1147.14	30.69
T_7^u	$Home_0$ → Central Park → East Ham Leisure Centre → $Home_0$	519	42	0.06	1122.18	3.70

by T_4, T_7, T_{11} and T_{17} not far behind with 0.92 of accuracy. Only for the travel routes T_{14}, T_{18} and T_{19} we achieve a very low accuracy meaning that such routes are present only in a very few number of the travel itineraries used as baseline. Accordingly to the result obtained for the location accuracy, the top travel routes are among the top popular locations.

3.2. Individual travel routes

In addition to collective mobility patterns, our methodology allows also to extract the behavior of individuals. In fact, some of the extracted frequent travel routes are traveled regularly by the same user or a very small group of users. In particular, we identified 102 users who produced a flow of tweets more regular, less sparse, presenting daily and monthly recurrences with an average number of tweets per user equal to 309. This value is higher than the tweets mean number per user of the collective trajectories, which is instead equal to 17. For each of those users we extracted at least one typical behavior. The extracted patterns represent London's residents routines and regularities.

In Table 7 we list the 20 top frequent travel routes of individuals, referred to as *residential travel routes*. In this case, it is not appropriate to rank the trajectories according to the popularity, because they represent individual behaviors. Notice that, those travel routes are traveled a large number of times by one or few users (at most 3) who repeat often the paths, in fact the number of journeys per users is high, consequently they are frequent but not popular. In this case, the entropy is low because visitors run through these routes in a periodic way, returning on different days. This periodic behavior could be typical in a work or educational place. For instance, T_1, T_3, T_9 and T_{10} involve work places. Focusing on T_1, it represents the movements from a railway station to a grocery store and vice versa. The users that periodically move between the two locations could be two employers of the store that arrive at work and come back home by train. T_1, T_{13}, T_{15}, T_{16} and T_{17} are characterized by visits to transport hubs; thus, they could be the routes of commuters who travel to work, home or reach a favorite shop or food place. Notice that, with the notation $Home_i$ we refer to the home location of a generic user i.

Due to space limits, in the following we present only one representative case of an individual (u) whose frequent travel routes are shown in Table 8.

The extracted mobility patterns show that user u moves daily from his home to a bar, which most likely is his workplace as evidenced by T_0^u, T_1^u and T_2^u, that represent the round-trip paths between those places. Looking at the other patterns of u

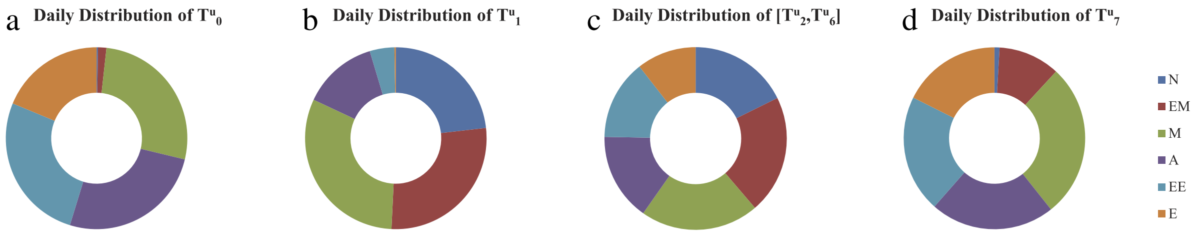


Fig. 10. Daily Distribution of Frequent Travel Routes, of user *u*.

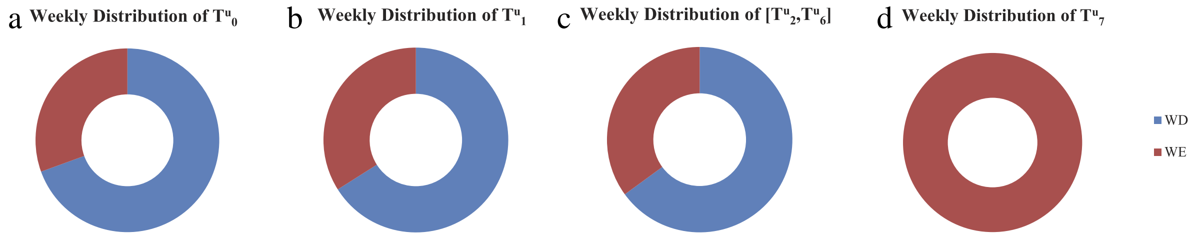


Fig. 11. Weekly distribution of frequent travel routes of user *u*.

in Table 8, one can note that as their length increases, their frequency decreases. Anyhow, they still present regularities in the mobility behavior allowing us to characterize *u*. For example, T_3^u, T_4^u and T_5^u show that the user goes to work using the metro; T_6^u allows to discover that the user goes to the pool before going to the bar; and T_7^u indicates that the user gets to the pool through a park before coming back home. As expected, the frequency of T_6^u and T_7^u is lower, probably because that user does not go to the pool and to the park every day of the week.

It is interesting to observe the graphs in Fig. 10 which highlight the time period of the day when the trajectories are typically traveled. Fig. 10(a) shows that the person travels along that route in different timeslots during the day, with a little prevalence the morning and early evening while he rarely follows the path during night and very limited times during early morning. This behavior could correspond to different profiles: (i) whether the person works all the day and returns home several times, for example for lunch, and then goes back to work; (ii) either she/he works during different timeslots in the days of the week. In this case, she/he could go back home in the morning because she/he worked during the night or could go back home in the afternoon as she/he worked in the morning. This is in accordance to the distribution of the opposite route shown in Fig. 10(b). In fact, from there we can see that the route home/bar is traveled mostly on early morning, morning and night.

Fig. 10(c) represents the average daily distribution of T_2^u, T_3^u, T_4^u, T_5^u and T_6^u. Those routes are represented together because their daily distributions are similar, meaning that the locations crossed by those trajectories are uniformly visited in different times of the day, with peaks during the early morning and the morning. Even T_7^u is traveled uniformly within the day except for the night where it presents a trough (Fig. 10(d)).

The graphs in Fig. 11 show the distribution of the routes between week days and week-end. Notice that all the trajectories, except for T_7^u, are traveled both during working days and week-end, with prevalence during the week days mostly for the travel route T_0^u. Fig. 11(d) highlights that T_7^u represents a typical behavior of the week-end, in which *u* probably spends his leisure time outdoor.

We can conclude that the proposed methodology allows to analyze the behavior of individuals, extracting their mobility patterns with associated timing information. This knowledge can be used to: (i) user profiling, for example, referring to the focused dataset, *u* could be a bartender; (ii) identify people preference and daily routines like traveling to work, visiting friends, recreation; (iii) activity prediction and recommendation; (iv) social ties.

Contrary to what was done for the collective travel routes, in this case we have no way of evaluating the accuracy of the individual mobility patterns.

4. Related work

The rapid rise of social networks has stimulated a lot of studies on human mobility and its implications in the social relationships and location-based services [9,15,16]. However, to the best of our knowledge, very few previous studies have been carried out on the extraction of trajectories by exploiting only the geo-references of tagged social posts. In fact, most of the related approaches in literature have been developed to identify recurrent patterns in mobility data originated by: (1) GPS traces of mobile devices; (2) geo-spatial and textual metadata associated to photos on web-like photo sharing sites.

With respect to the work done in trajectory pattern mining of GPS data [17,2,18–22], our approach presents elements of novelty that could be considered as an improvement in the field. The first major challenge is that we deal with the more unpredictable, sparse and irregular data emerging from location-based social networks, whereas GPS traces of mobile devices are highly available and sampled at regular time intervals. In particular, we deal with a large set of scattered mobility traces of geo-tagged data; most of the time such data locates a huge number of users only for a tiny fraction of the day. Therefore, these trajectories are usually generated at a low or an irregular frequency, leaving the routes between two consecutive points of a single trajectory uncertain. In addition, we work with trajectories at the semantic level instead of trajectories as raw points. Conversely, GPS data lack some semantics about the type of place in the travel sequences which would allow a better understanding of users' patterns. Similarly to the proposed work, Ref. [18] mines interesting locations and travel sequences in a given geospatial region. However, while we use a sequential pattern mining algorithm and a set of spatial–temporal features to extract and rank travel routes and locations, in [18] a HITS-based inference model is proposed. The work in [19] retrieves individual life patterns from raw GPS data, partially addressing the problem that we targeted as we mine both collective and individual mobility patterns. The works in [20–22] mine trajectories from GPS data but their aim is different from the one addressed by the proposed work. In particular, Ref. [20] mines the correlation between locations using individual's location history to achieve a personalized recommendation system, while Refs. [21,22] infer users' transportation mode, e.g., walking and driving.

Another group of works focuses on identifying hot spots and tourist routine behaviors from global collection of geo-referenced photos [5,16,23]. Photo-sharing sites, e.g., Flickr and Panoramio, contain billions of publicly accessible images taken virtually everywhere on earth. These photos are annotated with various forms of information including geo-spatial and textual metadata. In particular, our work shares the same aim as [5] of identifying the most frequent travel routes and the top interesting locations in a given geo-spatial region. However, the authors in [5] associate semantics to the locations on the basis of associated tags that are contributed by users for each photo on Flickr. In contrast, the geo-tagged tweets we analyzed are lacking of such information and we extract location semantics exploiting only the spatial–temporal information. Ref. [3] analyzes geo-tagged photos in Flickr and their semantic content to investigate: (i) the tourist movement patterns among regions of attractions (RoA) by exploiting Markov chain model, and (ii) the topological characteristics of tourist travel routes. Their focal point is analyzing tourist mobility at macro-level, while we also consider individual travel routes. Similarly, the authors in [7] focus on trip planning. They propose a travel route planning, Photo2Trip, which collects photos from Panoramio in order to suggest customized trip plans according to tourists' preferences.

Summarizing, the main difference with this branch of work is that it focuses primarily on images and leverages image features and tag-based data, whereas we work entirely only with spatio-temporal data. Moreover, deriving POI from geo-tagged photo datasets available on social media, could be inaccurate. Photos are usually placed at the point where they were taken and this point is often distant from the true POI location, thus, geo-tagging errors may occur. Given that, there is no control of user generated content, only a fraction of the photos tagged with a POI name actually depict it.

Most of the related works focusing on trajectory pattern mining of tagged social posts pursue a different goal from ours aiming at either: (i) identifying hot spots, i.e. areas of high density of movements, in an urban area to categorize regions through a coarse-grained mobility-based clustering; (ii) analyzing a set of pre-defined locations, the so-called POI [9,15,11]. Compared to the approaches of the first category we achieve a finer-grain detail in location detection allowing to identify precise venues in a urban area. The approaches based on POI (e.g., [11]) exclude from the analysis other important locations that are not classified as POI in the reference source (e.g., tourist guide) and from where many social data have been generated. Differently, we are able to identify all the locations where tweets have been actually posted by using a hierarchical clustering approach. Additionally, the works based on POI aim just at deriving itineraries from such pre-defined places, addressing, thus, just one of the step of our integrated methodology. For instance, in [24], the authors solve the problem of POI identification by mapping POI from Wikipedia to location tags of Flickr photos. Consequently, they provide personalized recommendations of only locations that in Wikipedia have been classified as POI excluding from the analysis interesting venues. Besides, differently from both the approaches, through our classifier we are also able to infer the semantics of a location and, thus, its category. In other words, given a place of which we know only its geographic coordinates and its spatial–temporal features we infer its category and extract relevant travel routes.

Other approaches studied mobility patterns in a city during exceptional events by observing micro-blog posts. As an example [25,26,10] used Twitter as a sensor to detect natural phenomena. However, compared to our proposal they perform a more coarse grain analysis of only typical trajectories by considering a fixed set of areas in a city. Also in this case, place semantics is not addressed.

Accordingly to the above discussion, the main contributions of our work are as follows:

- A first contribution is the integrated methodology which supports the overall trajectory pattern discovery process: (i) collection of raw tweets data and subsequent transformation into cleaned structured data; (ii) data-driven estimation of the parameters of the mining methods adopted; (iii) mining and analysis of the hidden information coming with the geo-tagged posts; (iv) synthesis of practical information in the form of popular travel itineraries.
- We effectively detect semantic locations from tweets. In particular, we are able to identify actually relevant locations through density-based clustering; then, we proposed a supervised classifier allowing to infer location semantics simply exploiting features specific to the locations and movements among them, like the duration of the stay, the number of visitors and the regularity of their behavior in the place.
- We define the problem of human mobility pattern mining formulating user movements as a sequence of visits to the detected semantic locations. In particular, for each user we compute a set of daily snapshots based on a set of novel concepts such as consecutive tweets, daily time slot and daily travel route.

5. Conclusion

In this paper we presented a novel methodology to extract and analyze the time- and geo-references associated with social data so as to mine information about human dynamics and behaviors within urban context. We performed a fine grain analysis of unpredictable and irregular information coming from geo-tagged tweets. In particular, we extracted a set of daily trajectories and we used a sequential pattern mining algorithm to discover frequent travel routes. We then, defined a set of spatial–temporal features over such routes and, accordingly, performed a statistical characterization of patterns, rules and regularities in moving trajectories. Future work include the integration of such methodology in recommender systems for trip planning, personalized navigation facilities, and location-based services useful for urban planning and management.

References

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

[2] E. Cesario, C. Comito, D. Talia, Towards a cloud-based framework for urban computing, the trajectory analysis case, in: CGC, 2013, pp. 16–23.

[3] Y.-T. Zheng, Z.-J. Zha, T.-S. Chua, Mining travel patterns from geotagged photos, ACM TIST 3 (3) (2012) 56:1–56:18.

[4] D. Villatoro, J. Serna, V. Rodríguez, M. Torrent-Moreno, The TweetBeat of the City: Microblogging used for discovering behavioural patterns during the MWC2012, in: CitiSens, Springer, Berlin, Heidelberg, 2013, pp. 43–56. (Chapter).

[5] Z. Yin, L. Cao, J. Han, J. Luo, T.S. Huang, Diversified trajectory pattern ranking in geo-tagged social media, in: SDM, 2011, pp. 980–991.

[6] T. Kurashima, T. Iwata, G. Irie, K. Fujimura, Travel route recommendation using geotags in photo sharing sites, in: CIKM, 2010, pp. 579–588.

[7] X. Lu, C. Wang, J.-M. Yang, Y. Pang, L. Zhang, Photo2trip: Generating travel routes from geo-tagged photos for trip planning, in: MM, 2010, pp. 143–152.

[8] V. Frias-Martinez, V. Soto, H. Hohwald, E. Frias-Martinez, Characterizing urban landscapes using geolocated tweets, in: SocialCom-PASSAT, 2012, pp. 239–248.

[9] J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh, Bridging the gap between physical location and online social networks, in: UbiComp, 2010, pp. 119–128.

[10] L. Gabrielli, S. Rinzivillo, F. Ronzano, D. Villatoro, From tweets to semantic trajectories: Mining anomalous urban mobility patterns, in: CitiSens, 2014, pp. 26–35.

[11] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, C. Yu, Automatic construction of travel itineraries using social breadcrumbs, in: HT, 2010, pp. 35–44.

[12] D. Falcone, C. Mascolo, C. Comito, D. Talia, J. Crowcroft, What is this place? inferring place categories through user patterns identification in geo-tagged tweets, in: MobiCASE, 2014, pp. 10–19.

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

[14] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth, in: ICDE, 2001, pp. 215–224.

[15] S. Wakamiya, R. Lee, K. Sumiya, Urban area characterization based on semantics of crowd activities in twitter, in: GeoS, 2011, pp. 108–123.

[16] D.J. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, Mapping the world's photos, in: WWW, 2009, pp. 761–770.

[17] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, in: KDD, 2007, pp. 330–339.

[18] Y. Zheng, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, in: WWW, 2009, pp. 791–800.

[19] Y. Zheng, X. Xie, Mining individual life pattern based on location history, in: MDM, 2009, pp. 1–10.

[20] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining correlation between locations using human location history, in: SIGSPATIAL, 2009, pp. 472–475.

[21] Y. Zheng, L. Liu, L. Wang, X. Xie, Learning transportation mode from raw gps data for geographic applications on the web, in: WWW, 2008, pp. 247–256.

[22] Y. Zheng, X. Xie, W.-Y. Ma, Understanding mobility based on gps data, in: Ubicomp, 2008, pp. 312–321.

[23] T. Rattenbury, N. Good, M. Naaman, Towards automatic extraction of event and place semantics from flickr tags, in: SIGIR, 2007, pp. 103–110.

[24] C. Lucchese, R. Perego, F. Silvestri, H. Vahabi, R. Venturini, How random walks can help tourism, in: ECIR, Vol. 7224, 2012, pp. 195–206.

[25] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors, in: WWW, 2010, pp. 851–860.

[26] T. Takahashi, S. Abe, N. Igata, Can twitter be an alternative of real-world sensors? in: HCI, 2011, pp. 240–249.