Mining Frequent Spatio-Temporal Patterns from Location Based Social Networks

Javier Béjar
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya (BarcelonaTech)
bejar@lsi.upc.edu

Abstract

Location Based Social Networks (LBSN) like Twitter or Instagram are a good source for user spatio-temporal behavior. These social network provide a low rate sampling of user's location information during large intervals of time that can be used to discover complex behaviors, including frequent routes, points of interest or unusual events. This information is important for different domains like route planning, touristic recommendation systems or city planning.

Other approaches have used the data from LSBN to categorize areas of a city depending on the categories of the places that people visit or to discover user behavioral patterns from their visits. The aim of this paper is to analyze the frequent spatio-temporal patterns that users share when visiting a city. This behavior is studied in a well limited geographical area by means of frequent itemsets algorithms in order to establish some causal dependence between visits that can be interpreted as interesting routes or spatio-temporal connections.

The data analyzed was obtained from the public data feeds of Twitter and Instagram inside the area of the cities of Barcelona and Milan for a period of several months. The analysis of these data shows that these kind of algorithms can be successfully applied to data from any city (or general area) to discover useful patterns that can be interpreted on terms of the city singular places and areas and that these patters can be used as a the elements of a knowledge base for different applications.

Keywords: Location Based, Social Networks, Frequent Itemsets, Smart Cities, User Profiles

1 Introduction

Location Based Social Networks [15], like Twitter or Instagram, are an interesting source for user geospatial and temporal behavior analysis. The data that these applications provide include user's spatio-temporal information that can be used to uncover complex behaviors and patterns, including frequent routes, points of interest, group profiles or unusual events. This information is important for applications such as city management and planning or recommender systems.

The goal of this paper is to analyze these spatio-temporal data using frequent itemsets algorithms in order to find out what patterns arise from user behavior in large cities. The data used in the experiments was obtained from Twitter and Instagram social networks in the geographical area that surrounds the city of Barcelona and Milan. The expectation is that this unsupervised technique could be applied to any city (or general area) in order to discover useful patterns. The aim is that these spatio-temporal patterns could be used later for application domains that need more structured information about a city for reasoning and making decisions about the activities of citizens.

This study is included inside the European ICT project SUPERHUB, that has among its goals to integrate different sources of information to help and improve the decision making process oriented to improve urban mobility. This project is part of the EU initiative towards the development of smart cities technologies. Two of the key points of the project are to use the citizens as a network of distributed sensors that gather information about city mobility conditions and to generate mobility

profiles from these users. This information will be used to implement route planning and mobility recommendation systems.

The plan of the paper is as follows: Section 2 introduces to other approaches to discover patterns from spatio-temporal data in general and from LBSN in particular. Section 3 describes the characteristics of the data used in the experiments and the transformations applied to obtain datasets suitable for applying the unsupervised data mining algorithms. Section 4 explains the approach, by means of frequent itemsets algorithms, to discover frequent patterns as an approximation to the common regions of interest of the users and their connections in the city. Section 5 shows the results obtained of applying the described techniques to the data collected for Barcelona and Milan from Twitter and Instagram. Finally in section 7 conclusions about the results of the different techniques are explained along with the possible extensions of this work.

2 Related work

Since the wide availability of devices capable of transmitting information about the location of users (mobiles, GPS devices, tablets, laptops), there has been an increasing interest in studying user mobility patterns inside a geographical area. These data are available from different sources, like GPS traces extracted from these devices or from internet sites where users voluntarily share their location among other information.

Different knowledge can be extracted from these data depending on the analysis goal. One important application is the generation of visualizations, so patterns in the data can be easily identified and interpreted by experts in the specific domain of analysis, for example, city officials studying citizen mobility and traffic distribution. In this line of work, [2, 3] describe different methods for obtaining visualizations of clusters of GPS trajectories extracted from the movements of cars inside the city of Milan. In [11], different techniques based on Kernel Density Estimation are employed to detect and visualize hot spots in the domains of epidemiology and criminology.

Other applications include user routine mining and prediction. The idea is either to recognize user activities from the repeating temporal behavior of individuals or groups, or to recommend to users activities according to past behavior or user context. Data gathered from mobile phones of MIT students and faculty was used in [5] to predict user routines and their social network using hidden markov models and gaussian mixtures. The same dataset was used in [6] for user and group routine prediction, user profiling and change discovery in user routines. The applied methodology used a text mining analogy, considering individual activities as words and sequences of activities as documents. This allowed to transform the user activities to a bag of words representation and then to cluster them using Latent Dirichlet Allocation. In [14], data collected from GPS trajectories over an extended period of time inside a city is used. From these trajectories, a set of special points were extracted, defined as points where a user stays in a bounded region for a short period of time. These were considered as points of interest for the user. These points and the categories of the places inside the region surrounding the points of interest were used as the base for a touristic recommender system.

An alternative to information collected by GPS enabled devices are the Location Based Social Networks. The main issue with GPS data is that it is difficult to obtain this information continuously for a large number of users. These social networks allow to sample information from more users, with the drawback of having a much lower sample rate. The previously mentioned text mining analogy is used in [8] to analyze data from Foursquare. Only information relative to the category of the check-in places was used, and all the check-ins of a user were put together to represent his global activity. Latent Dirichlet Allocation was then applied to obtain clusters described by sets of salient activities. These sets of activities allowed to characterize the different groups of persons in a city as a first step to extract user profiles to be used for different applications. Data from the BrightKite social network (similar to Foursquare) was analyzed in [10] to obtain geographical profiles of the users and to measure the correlation between their activities and their geographical locations. In [13], data from Twitter was used to predict user activity. From the collected Twitter events, only the ones corresponding to Foursquare check-ins were extracted. Different clusterings of the events were

obtained using as characteristics spatial location, time of the day and venue type. These clusters were used as characteristics for activity prediction and recognition. The venue types were transformed to a set of predefined activities (lunch, work, nightlife, ...) and the prototypes of the clusters were used to obtain activity predictions for new data.

More close to the analysis presented in this paper, in [9] almost three months of Foursquare check-in data were extracted from tweets inside the area of Japan. These data were geographically clustered using the EM algorithm. The daily behavior of groups inside the cluster was represented by dividing the day in four periods, computing different types of counts of the events and using as attributes the sign of the difference of the counts between consecutive periods. With these attributes a database of transactions was generated for all the days in each cluster. A frequent itemset algorithm was used then to discover the most frequent behavior patterns for each cluster. An a posteriori analysis of the venues that the clusters contained, given their frequent patterns, allowed to characterize the clusters depending on the distribution of the categories of the venues.

3 The dataset

The aim of this paper is to show the feasibility of extracting meaningful relational patterns from LBSN for city analysis, so our interest was focused on the data that can be obtained from the most popular of these kind of social networks, specifically Twitter and Instagram.

The data used in the experiments was collected from the public feeds from both social networks during a period of several months. A priori, the quality of these feeds can be considered as non optimal due to the availability limitations that are imposed by these social networks to free data access. For example, the Twitter public feed (Twitter Streaming) provides a random sample with a size that has a maximum of a 1% of the total number of tweets at each moment. There has been some studies of the quality of this specific data source [12] that show that the sampling provided is biased and can be misleading depending on the type of analysis. Although, these studies also point out that it is possible to identify around 50% of the key users on a given day, accuracy that can be increased when the period of collected data is large. This period is several months in our case. Also, given that the data is collected using a bounding box around a geographical area, the small percentage of events obtained will proportionally account for a larger percentage of the total tweets inside that area. Due that we are interested in the frequent activity of users, and that the period of data collection is large enough, we consider that the data are representative enough to provide meaningful results.

All the events obtained from these applications (tweets, photographs) include spatio-temporal information represented as latitude and longitude, geohashing and timestamp. A unique user identifier is also provided for each event that allows to relate all the events of a user. With this information we can obtain a low rate sampling of the spatio-temporal behavior of a large number of users.

As mentioned, the data obtained from Twitter and Instagram was geographically constrained. The events were filtered to extract only the ones approximately inside an area of 30×30 km^2 of the cities of Barcelona and Milan. The size of the area was chosen to include all the populated areas of the cities and other surrounding cities, so also the behavior of the citizens in these cities were included, allowing this way to extract mobility patterns outside these cities and not only internal mobility.

The dataset for Barcelona was collected during an eleven months period (october 2013 to august 2014), the number of events extracted from Twitter and Instagram is around three millions each. The dataset for Milan was collected during a six months period (march 2014 to august 2014), with around a million Twitter events and half a million Instagram events. The difference in the sizes of the datasets, apart from the period of data collection, is accounted by the larger number of tourists that visit Barcelona. The population inside the area selected for Milan is actually larger than the population for Barcelona (7.4 millions versus 5.3 millions).

Despite the large number of events data are actually very sparse from the user-event perspective. Both datasets present similar user-event distributions, where around 50% of users only generate one event on a given day and 40% of the users only generate one event during the collected period. In the next subsection a more detailed statistical analysis of the dataset is performed.

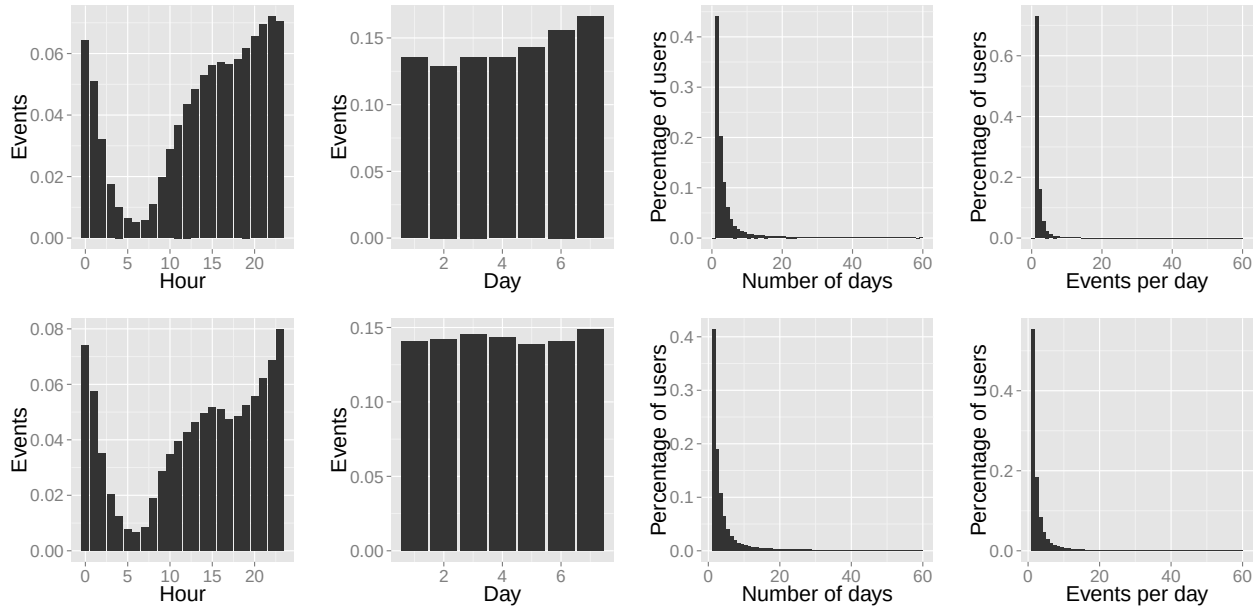

Fig. 1: Hourly and dayweek events, distribution of number of days per user and distribution of the number of events per day (Up: Barcelona - Instagram / Down: Barcelona - Twitter)

3.1 Descriptive statistical analysis

In order to understand the characteristics of the data some simple descriptive statistical analysis was performed including frequency of events for natural periods like weeks and hours. Also the percentage of users according to the number of events generated and the number of days that have events during the period of data collection was analyzed.

The figure 1 shows different information of the events from the data collected for the city of Barcelona. The first plot of the figures represents the hourly percentage of events. Both social networks show a similar tendency, the distributions seem to show two modalities, one centered around 14-15 hours and another centered around 22 hours. This shows that there are two distinct people behaviors, one that generates events during the day and other during night hours. This tendency is more clear on the Twitter plot, showing a decrease of activity at 16-17 hours that marks the beginning and the end of these behaviors. A reasonable explanation is the daily cycle work-leisure, during work hours there is less people with the time to generate events and during leisure time they have more time and also more events that they want to publish.

The second plot shows the weekly distribution of the events. It can be seen that Instagram is more used during the weekends. A possible explanation in that, given that it is a photo sharing social network, it is more probable to have something to show during weekends than during working days. This tendency does not appear in the Twitter data, probably because it takes less time to write an small text than to take a photograph.

The third plot is the percentage of users respect to the number of days they have any activity during the collected period. Given that the data is a small random sample of the actual events, the probability of capturing repeatedly a casual user in several different days is very small. Also, the large number of tourists that visit Barcelona makes that a significant number of users only generate events for a small period of time. This explains that more than 40% of the users only have one day of events during all the period. The distribution of the percentage decreases exponentially with the number of days (following a power law), being the percentage of users captured in more than 10 different days very small. It has to be noticed that both networks show the same distribution.

The fourth plot is the distribution of the daily number of events per user, that is, how many events usually a user generates in a given day. Because only a random subset of the events is captured, most of the users will have a number of events very low in a day, so this graph gives just an idea

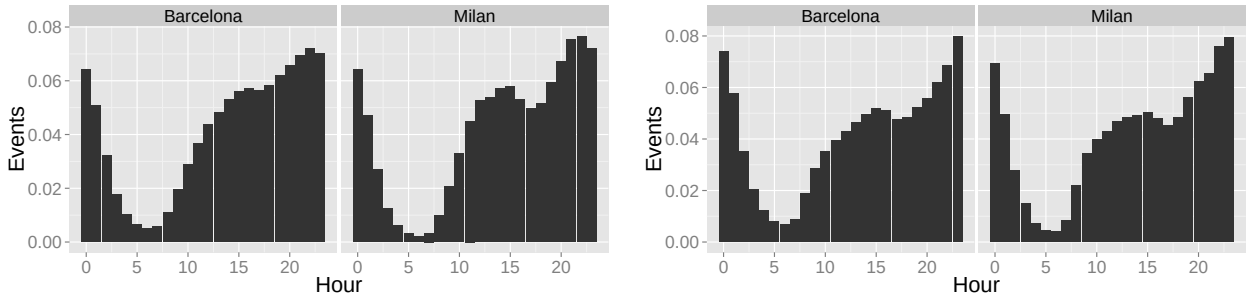

Fig. 2: Hourly events for Instagram (left) and Twitter (right) comparing Barcelona and Milan distributions.

of the statistical distribution. The actual distribution should decrease more slowly with the number of events. The parameters of the distributions for both social networks are slightly different, but following the same power law. For Instagram almost 70% of the users generate only one event in a day, being around 50% for Twitter. The reason could be also that it is easier and faster to write a text that to post a photograph. The distribution of the number of events also drops faster for the Instagram data, being almost negligible the number of users that generate more than 4 events in a day.

The graphics for the city of Milan are fairly similar, there are slightly differences due to the circumstance that Barcelona receives a larger number of tourists. As an example, the figure 2 shows the comparison between the hourly number of events for the two social networks for both cities. For Twitter the behavior is almost identical in both cases, but for Instagram there is a slightly difference between them. The cycle work/leisure in the city of Milan shows a decrease of events like for all the Twitter graphics. This means that there is a lower (or no) influence of city visitors in that cycle, making the behavior of Instagram and Twitter more comparable for this case.

4 Discovering spatio-temporal patterns

In order to study the patterns in the data, we have to define some discovery goals. Our main purpose is to reveal relationships among the events that the users generate in a global manner, in order to discover general patterns that will be suitable for being applied to different domains. Frequent individual events can be studied as a first step of analysis but the can not reveal more complex behavior than popularity, and are not strictly interesting for many application domains. To include the temporal dimension allows for a more complex analysis by adding causal information.

Henceforth, the goal for the analysis of the dataset is to discover connections between geographical places. We consider the individual events as the building blocks of the patterns and the connections in time among the events as the patterns. The assumption is that if the same set of events, considering an event as being in a geographical place around certain interval of time, is generated by a large number of people, then it is probably significant. These patterns can be useful for example to discover places that are visited frequently by people that have a certain profile (tourists), places in the city that need to be connected by public transportation or traffic bottlenecks that are connected in time.

This goal is similar to market basket analysis. In this type of analysis a set of transactions from the purchases of users in a visit to a store are defined. The patterns discovered are associations among products defined by the frequent purchase of items by groups of clients. In order to transform our goal into this terms, we have to define what is a transaction in our domain. We can define, in a similar way, a transaction as the daily activity of a user, being the activities the geographical places and the time of the events. This will allow us to apply the same techniques used by market basket analysis to discover spatio-temporal points that are frequently visited by groups of people at similar times, representing their association.

In the next subsections we will explain how to define the transactions from the LBSN data and the techniques used for the patterns extraction.

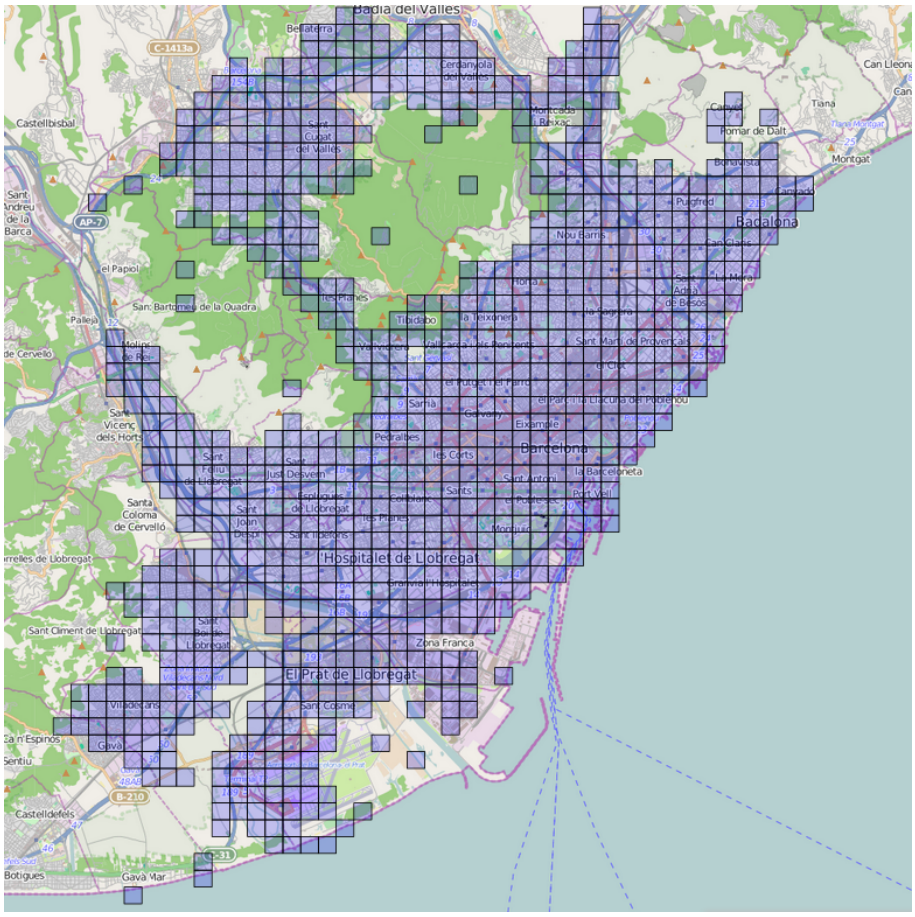

Fig. 3: Grid of Barcelona Twitter events (grid size 500m, more than 25 tweets)

4.1 Defining transactions

It is difficult to extract patterns of the behavior of the users directly from the raw events, so, first the data has to be transformed to a more useful representation. There are different ways of performing such transformation. It was considered that the behavior of a user during a day may contain useful information for the application domains, so it was decided to generate transactions from the users daily events. The attributes of such transactions are the presence/absence of an event in a geographical position (latitude, longitude) during an specific range of hours of the day.

The geographical positions correspond to point coordinates inside the cities. Being the purpose to use the frequency of the events as the basis of the discovery, it is obvious that the raw coordinates can not be used directly. The probability of having a large number of events at the same coordinates is extremely low. This means that the geographical positions have to be discretized in some way to increase the probability of coincidence of the events. The same problem appears for the temporal dimension.

Different discretizations of a geographical area can be proposed to allow the extraction of patterns at different resolutions and complexity. The first approach that was used in this analysis is to divide the geographical area in a regular grid. Usually, not all places in a geographical area can be accessed, so the actual number of places a user can be is much less than all the possible cells in the grid. For example, for the data from Barcelona, only around half of the grid cells have at least one event. The use of a grid also allows to study the events at different levels of granularity. Controlling the size and shape of the cells it can be defined a discretization that captures from specific places (e.g. points of interest) to organizationally meaningful regions (e.g. city districts).

The main advantage of this method is that the cost of grouping the events is linear respect to the number of events. The main drawback is that some frequent events could be lost if their counts are divided into adjacent cells by the discretization. A larger granularity can reduce this problem with

the cost of the loss of information. Figure 3 shows the cells for Twitter events with more than a given threshold during all the recording for a grid with size cell of $500m \times 500m$ for the city of Barcelona during the data collection period. It can be seen that there are events present in all the inhabited places of the selected area and the main access roads that connect Barcelona and its surrounding cities.

Given that a regular grid cannot adapt to the different geographical densities of the events, to use a clustering algorithms to generate a discretization that discovers these densities could be an alternative. From all the clustering algorithms that can be applied, an interesting possibility is to use an incremental clustering algorithm in order to be able to update the model with new events and even to adapt the model with changes in the behavior of the data.

So, as a second data transformation, the leader clustering algorithm [4] (see algorithm 1) is used to group the spatial coordinates of the events. This algorithm obtains spherical clusters grouping incrementally examples that are inside a predefined radius. This radius has the same effect than the size of the grid, obtaining thus different dicretization granularity with different values. The main advantages are that the clusters adapt to the density of the data and that it also can be computed in linear time respect to the number of events. Figure 4 shows the clusters obtained by clustering the events with a diameter of $500m$. It can be seen that the centroids of the clusters do not fall in a regular pattern, adapting to the different densities of the events and reducing the possibility of splitting close events in several clusters.

Another possibility would be to use a density based or grid based clustering algorithms to extract dense areas from the raw data. The computational cost would depend on the specific clustering algorithm, but being it usually quadratic in the number of examples and considering that this a very large dataset, it could arise scalability issues.

The discretization of time is a more simple issue. Two possibilities arise, the first one is to define a set of same size hour intervals and assign to the events a timestamp according to these intervals. The second one is to use the hourly distribution of the events during the day to find a discretization meaningful for the data. As we have shown in 3.1, there are two modalities in this distribution with means around 16 and 21 hours respectively. These two modalities allow to split the day interval in different ways depending on the number of intervals. It also has to be noted in the data distributions that a day begins and ends around 6 am. This means that a daily transaction has to include events inside this hours interval to be correct.

Using these transformations we can build a transactions dataset defined by daily events grouped by a geographical area (defined by the geographical discretization method) and by their timestamp (defined by the time discretization method). To this transformed dataset we can apply frequent itemsets/association rules algorithms to uncover the patterns in the data.

4.2 Generating frequent routes

In our formulation, a transaction contains events that correspond to geographical areas during different ranges of time. It can be considered significant/interesting to observe a number of simultaneous events higher than a specific threshold. This events could be linked by temporal and/or causal relations that have to be interpreted by a domain expert. A natural interpretation is to consider their relations as routes or connections between different geographical points at different points of time.

The extraction of these sets of frequent events can be obtained by the application of frequent itemsets algorithms. The main issue with mining frequent itemsets is that all the possible subsets of items have to be explored to determine their frequency. For our problem, the number of possible itemsets is very large, even for very coarse discretizations. For example, for the Barcelona dataset, using a grid with cells of $500m$ side length, it means to have transaction with 3600 geographical positions multiplied by the size of the time interval discretization. To use the clustering dicretization would reduce this number, depending on the geographical distribution of the events, but within the same order of magnitude.

Frequent itemsets algorithms reduce this computational problem by exploring the itemsets ordered by their size and taking advantage of the anti monotonic property of support. Namely, an itemset

Algorithm 1 Leader Clustering Algorithm

```
Procedure: LeaderClustering(radius,Data)
Centroids[0] = Data[0];
nexamples[0]= 1 ;
nclusters = 1;
for i = 1..n do
    ncen, dist = NearestCentroid(Centroids, Data[i]);
    if dist ≤ radius then
        Centroids[ncen] += Data[i];
        nexamples[ncen] += 1;
    else
        Centroids[nclusters] = Data[i];
        nexamples[nclusters] = 1;
        nclusters += 1;
    end
    return (Centroid/nexamples)
end
```

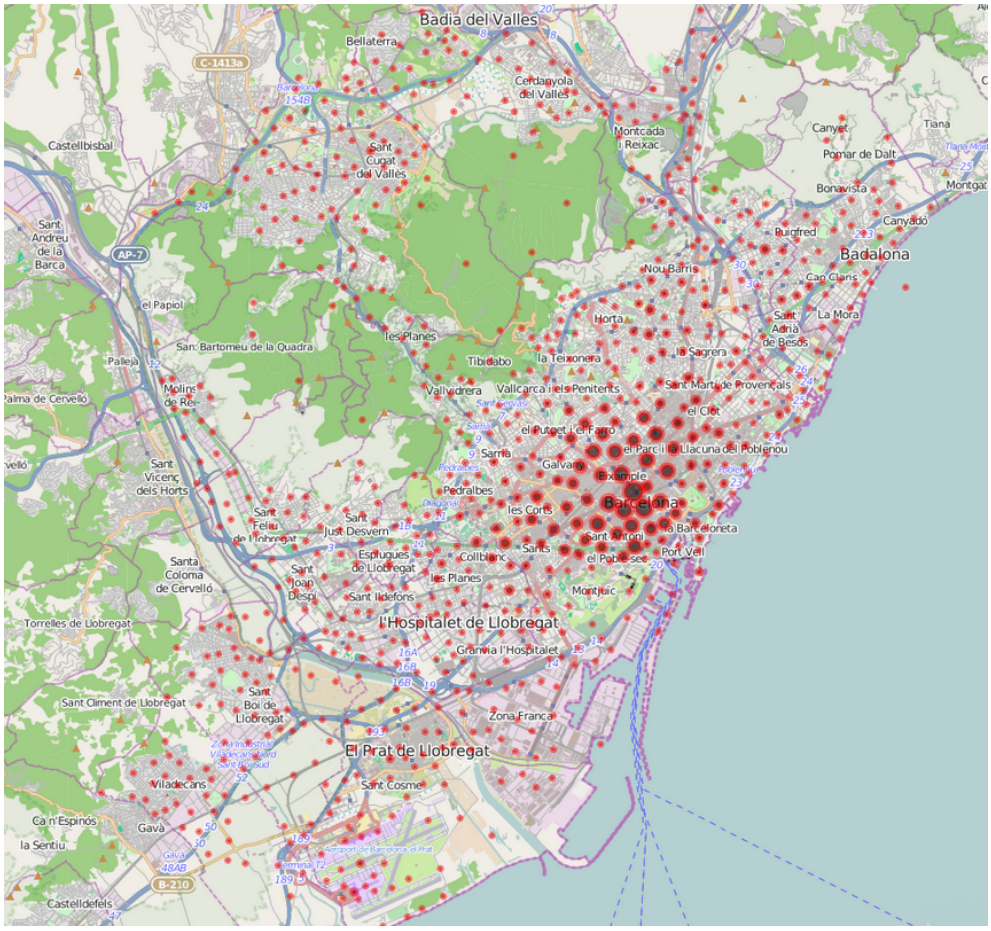

Fig. 4: Clustering of Barcelona Twitter events (radius=500m, more than 25 tweets) the size of the clusters is proportional to the number of events in each cluster

only can be frequent if all its subsets that contain one element less are also frequent. There are different algorithms for frequent itemsets discovery with different approaches. The classical Apriori algorithm [1] and related variations would not be suitable for this problem given the large number of different items that a transaction has. A better approach is the one proposed by the FP-growth algorithm [7]. This algorithm avoids to generate possible candidates by summarizing the transactions of the database using a specialized data structure called FP-tree. This data structure uses a prefix tree approach to store the transactions as strings, assuming an arbitrary lexicographical order for its items. The frequency of every item is summarized in the nodes of the data structure according to the count of prefixes each item shares with other items.

The algorithm for extracting the frequent itemsets uses this structure to perform the exploration. It begins with all items that appear on the first level of the FP-tree, generating all possible frequent itemsets that begin with that item and have a frequency larger than the defined support. Then, it continues traversing recursively through the data structure adding more levels to the itemsets until no more frequent items can be included. The algorithm uses a divide and conquer strategy, so for each item that can be added to the itemset, the FP-tree is divided and explored in parallel if necessary. This circumstance makes the algorithm very scalable, being able to process datasets with large numbers of items and transactions (see [7] for the details).

Applying these algorithms different subsets of the total set of frequent itemsets can be extracted. Obtaining all possible itemsets that have a support larger than a threshold usually results in a very large number of redundant patterns that are difficult to interpret. In our case these redundant patterns will represent all sub routes of a larger route. Another possibility is to only extract the *closed itemsets*, being those the ones where a change in the value of the support appears respect to the next level. This constraint can also results in redundant itemsets. For the applications that we are interested in, the more suitable subset of itemsets is the one that contains only the longest possible frequent itemsets, that kind of patterns are called *maximal itemsets*. These patterns are the largest itemsets that have a support larger than the minimum support threshold. These patterns will be the ones used to summarize all the connections between places and time slots and will represent the wider set of points that are frequently connected during a day.

One issue for these algorithms is to select an adequate value for the support parameter. Given the sparsity of the data, the selection of this minimum support is a difficult problem. The domain indicates that for a pattern to be significant, the number of people that presents it has to be large. Although, given that we have an incomplete picture of the users behavior, because of the random sampling, the actual value has not to be so strict that no patterns are generated.

It also has to be considered that the density of events is different in different parts of the city. As can be seen the figure 4, that represents the clusters proportionally to their number of events for the Barcelona Twitter dataset, the center of the city concentrates a large part of the events. This would suggest that using a larger support will results in only patterns in this part of the city. A lower support could also multiply the connections between the center of the city and the rest. In our experiments we will consider different support thresholds to explore how the number and characteristics of the routes change.

5 Datasets analysis

In this section we explore the different parameters of our proposed approach. The grid and clustering discretizations will be also compared in order to determine the more adequate technique for transforming the event data on transactions.

For the granularity of the grid/clustering, different values will allow to examine the patterns at different levels of abstraction. Taking in account the sparsity of the data, we consider that reasonable values for this parameter should be between 500 and 100 meters. A lower value will generate cells/clusters with a very low number of events that will result in no frequent patterns.

For the time discretization, the events show two distinct hourly distributions during the day (see figure 2). A discretization following these distributions seems the more reasonable choice. This means

that a day actually begins and ends around 6am. A day can be discretized for example using the following intervals:

- A discretization in two ranges, one beginning at 6am and ending at 6pm and another beginning at 6pm and ending at 6am of the next day. This accounts for the two distinctive populations of events.
- A discretization in three ranges, one beginning at 6am and ending at 4pm, another beginning at 4pm and ending at 10pm and another beginning at 10pm and ending at 6am of the next day. This accounts for the two distinctive populations of events, but separating the range where the populations are mixed.
- A discretization in four ranges that splits each interval of the first discretization in two at the mean value of the distributions, namely 4pm and 10pm.

As mentioned before, frequent itemsets algorithms use a threshold as significance measure. This value has to be defined experimentally guided by the knowledge domain. Considering that a large of users only generate a event per day (see 3.1) and that the number of events decreases following a power law, it is desirable for the support parameter not be too high if we want any pattern to appear. Also, if we want to discover connections among places that are not only the popular ones a low value for the support is necessary.

For the experiments we will consider reasonable to obtain patterns that at least include 25 events and the effect of increasing the value of this support will be studied.

5.1 Grid discretization vs clustering discretization

First we will experiment with the two proposed discretization methods. As mentioned before, the main possible issue with grid discretization is that the grid could split the event densities into several adjacent cells resulting in the inability of discovering some patterns. This discretization can also result in the discovery of spurious patterns between adjacent cells and duplicated patterns relating two or more adjacent cells with the same cell. Using a small size grid (100m) these problems effectively appear in the patterns discovered. This problem could also appear with the clustering discretization that uses the leader algorithm, but the number of cases is much lower. Figure 5 shows a comparison between the patterns found using grid discretization and clustering discretization. It can be appreciated that several spurious and split patterns appear for the grid discretization and this circumstance is appreciably reduced with the clustering discretization.

Another problem presented by the grid discretization is that the representation of the patterns in a map looks less natural. The routes are translated to the map using the centroid of the cells that usually do not correspond to a natural geographical point in the city (a street for example). Another issue is that patterns that are closer appear as horizontal or vertical lines in the representation, degenerating to a grid when the size of the cells is large. With the discretizations obtained using clustering, the centroids of the clusters usually correspond to natural points given that are the areas with larger density of events and the representation seems more natural even with a coarse discretization (see figure 6).

Given these results, for the rest of the paper only the clustering discretization will be used.

5.2 Algorithms parameters

The support threshold determines what patterns are significant in the dataset. Given that the users events are a sample of the total number of the actual events and that a large number of users appear in the dataset only for a short period of time it is reasonable to use a low support threshold.

This value also depends on the specific behavior of the users of the LBSN inside the geographical area. For example, if the events are generated mainly by people visiting the city, a large number of events will be concentrated at specific points, allowing for a higher threshold. Although, if the events

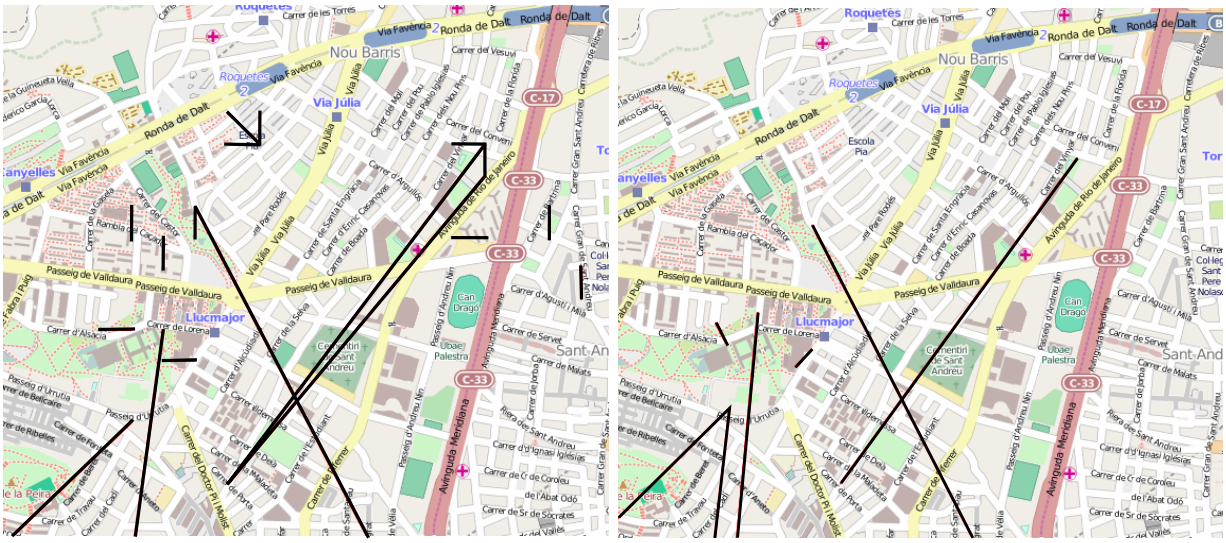

Fig. 5: Routes extracted using grid discretization (left) versus clustering discretization (right) with 100m resolution

Fig. 6: Routes extracted using grid discretization (left) versus clustering discretization (right) using a coarse discretization of 500m resolution

correspond mainly to people living in the city, the events will be more distributed geographically. This shows in the experiments performed. As can be seen in table 1, for the data from Instagram in Barcelona, the number of patterns is larger than for the Twitter data, even considering that the number of events is almost the same. The difference is even larger when the discretization is finer, only explained assuming that the patterns are geographically more concentrated for the Instagram dataset.

The ratio of the decreasing of patterns when the support is increased is similar for both social networks and for both cities. Because the percentage of users with a large number of events in a day decreases following a power law with similar parameters for all datasets, it is expected that the decreasing of patterns with the increase of the support also follows this behavior.

In the same tables, it can be observed the effect of the time discretization. For the Barcelona datasets, using a larger number of intervals reduces the number of patterns, this also happens for the Instagram data for Milan. This decrease seems to be proportional to the density of events in the time intervals. This effect is greatly reduced in the Milan Twitter data (see table 2). It seems that the events are more homogeneous and concentrated in the different intervals for this dataset and the chance of their density being broken by the discretization is lower. For the different domain

	Time	Two Intervals			Three Intervals			Four Intervals		
	Diameter	100	250	500	100	250	500	100	250	500
Support	25	567	2076	3368	408	1557	2870	332	1153	2427
	50	114	593	1174	71	352	850	61	260	635
	100	24	136	356	16	79	217	14	50	138

	Time	Two Intervals			Three Intervals			Four Intervals		
	Diameter	100	250	500	100	250	500	100	250	500
Support	25	1014	3075	3680	675	2376	3306	497	2015	2878
	50	267	1108	1545	141	817	1241	95	595	1075
	100	59	370	604	28	243	463	21	179	369

Tab. 1: Number of patterns generated for different values of time intervals, clustering diameter and support for the Barcelona Twitter (up) Instagram (down) data

	Time	Two Intervals			Three Intervals			Four Intervals		
	Diameter	100	250	500	100	250	500	100	250	500
Support	25	217	370	713	210	315	608	268	346	596
	50	31	78	184	59	80	155	80	97	147
	100	13	25	44	17	24	43	17	21	34

Tab. 2: Number of patterns generated for different values of time intervals, clustering diameter and support for the Milan Twitter data

applications, it could make more sense to have a larger number of intervals for interpretability reasons, but different kinds of interpretations can be extracted from all the proposed time discretizations.

For the diameter of the cluster discretization, also as expected, to reduce the granularity reduces the number of patterns. The proportions differ with the support and the time discretization. An explanation could be that when the support is high only a small subset of users are considered and their events will be more disperse geographically, so a larger granularity is needed for grouping them.

In the following section we will explore the results obtained by using the extreme values in the ranges of these parameters for both datasets.

6 LBSN patterns

All the experiments have been performed using the events from the 70.000 more active users in the datatasets. By active meaning the users that have several events during all the period or some events concentrated on a small time period. For the Barcelona Twitter dataset the 50 more frequent users were discarded because they corresponded to different automatic bots, like traffic conditions, weather reports or spammers. The number of transactions generated for each dataset is over eight hundred thousand for Barcelona Twitter, over one million two hundred thousand for Barcelona Instagram, over two hundred fifty thousand for Milan Twitter and around three hundred thousand for Milan Instagram. Each transaction corresponds, as explained before, to the events generated by a user during a day discretized using the leader clustering algorithm and using an specific time discretization.

Analyzing the data, results from the patterns obtained from Twitter and Instagram show different perspectives on the city. For example, the most frequent Instagram routes are generally related to tourist activity. Many of the frequently related areas show connections among touristic points of interest. For example, for the city of Barcelona the Sagrada Familia, Plaça Catalunya, Güell Park, Casa Batlló and La Pedrera appear among the most connected points of interest, and the rest are concentrated around the center of the city. Patterns from Twitter events are more diverse and show behaviors also for the actual dwellers of the city. This shows in the diversity of places that appear

connected, that include touristic points of interest but also other places distributed all over the city. As an example of patterns for Barcelona, some of them show people from the nearby cities that tweet at the beginning or the end of the day from home and also arriving to their place of work inside the city, or people that traverse the city from their home outside Barcelona to their workplace in another city near Barcelona.

More in detail, for the Barcelona Instagram data, with the most constrained parameters, using a 100m discretization, morning/evening time discretization and extracting only patterns with a support of more than 100 events, 59 routes are obtained (see figure 7, left) that represent the most photographed places in the city and how are connected. All routes have length two, this is expected because the support is high. Some routes appear several times with different time relations, having for example routes that connect two places in the morning or one in the morning and another in the evening. Twitter for the same parameters reduces the number of routes to 13, it shows also tourist activity (figure 7, right), but also other connections appear related with people in areas with high nightlife activity. In this set of routes there are four with length three, some connect nearby places, and other correspond to return routes.

For these parameter the Milan Instagram generates just two patterns around the Duomo cathedral in the center of Milan. Milan Twitter data generates 13 patterns (see figure 8), this patterns are longer that for Barcelona, only six patterns of length two, the rest correspond to patterns of length three (1), four (5) and five (1). The longer routes correspond to return routes, the same points in the morning repeat in the evening and are situated near the main routes in the center of the city, probably indicating traffic jams at rush hours.

When the parameters used to extract the patters are relaxed, a more complete view of the cities is obtained. Using a coarser granularity with a 500m diameter, with four intervals time discretization and a support of 100 events Barcelona Instagram obtains 369 patterns, mostly of length two, that expand around the center of the city and also include other significant areas outside the center (see figure 9, up). Most of the identified routes are still related to touristic points of interest. Barcelona twitter data increases the patterns outside the touristic points of interest even when the number of routes is significantly lower, 139 in this case. It is interesting (even when expected) that chains of patterns appear at the main entrances of the city and around the main train station, suggesting rush hour patterns. Also appears the connection between Barcelona El Prat Airport and the center of the city. There are several public transportation connections between the city and the airport, but with this level of support it is the only one that appears. This means that a large number of users prefer the dedicated bus line that connects the airport with the city center over all other alternatives. This alternative is probably preferred by tourists arriving Barcelona because they have less knowledge about the other possible public transportation alternatives.

For these parameters, Milan Instagram data does not extract any patterns, meanwhile for Twitter data 34 patterns are found. These patterns expand the connections between possible rush hour points around the center of the city and also include a connection with a touristic area that contains a castle (Sforza castle) and several art museums.

Patterns obtained with coarser discretization and lower support also reveal the less known part of the cities and their metropolitan areas. When support is reduced to 50 events, for Barcelona Twitter data, the connections with the cities inside the Barcelona metropolitan area begin to appear. These connections can be mapped to the different public transportation alternatives from these cities to Barcelona. A support level of 25 events, not only increases the connections with these cities and Barcelona but also discovers connections among these cities and patterns inside these cities (see figure 10 for an example). Connections from the airport also increase, mainly including connections with subway, train and bus stations. One of these connections is with the FC Barcelona football stadium, meaning that there are people that visits the city only for football matches. Barcelona Instagram data using less support includes other points of interest that are of less interest for the common tourist or that are visited by people that stay in the city longer, like the FC Barcelona stadium or the Tibidabo Park.

For Milan, Instagram data with a support of 25 events includes also other touristic related areas

Fig. 7: Frequent connected areas from Barcelona Instagram (left) and Twitter (right) data (Leader Clustering 100 meters, two time intervals, support = 100)

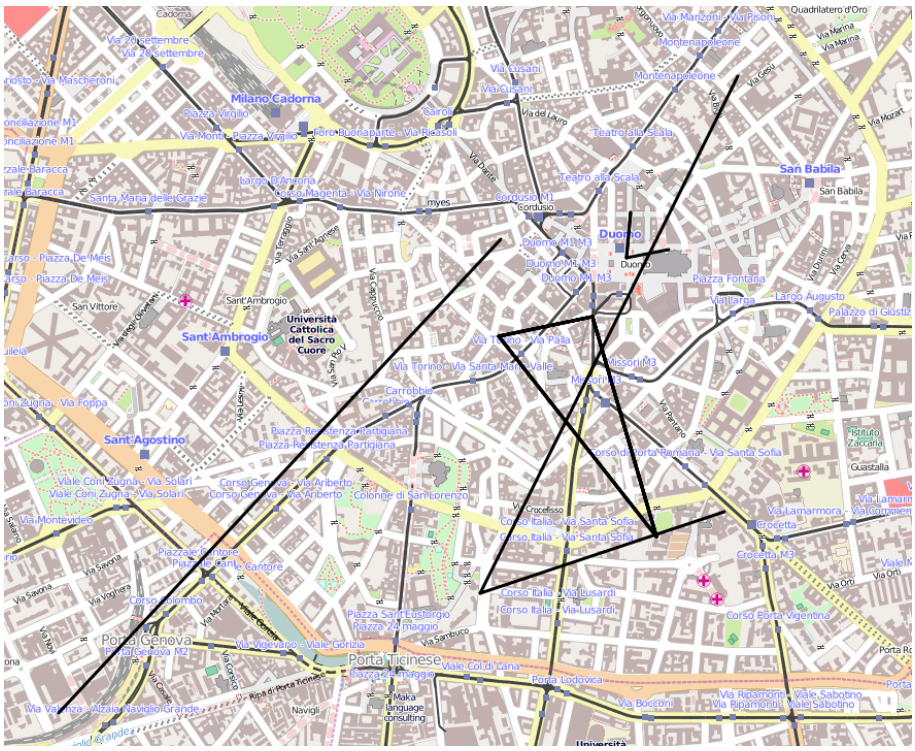

Fig. 8: Frequent connected areas from Milan Twitter data (Leader Clustering 100 meters, two time intervals, support = 100)

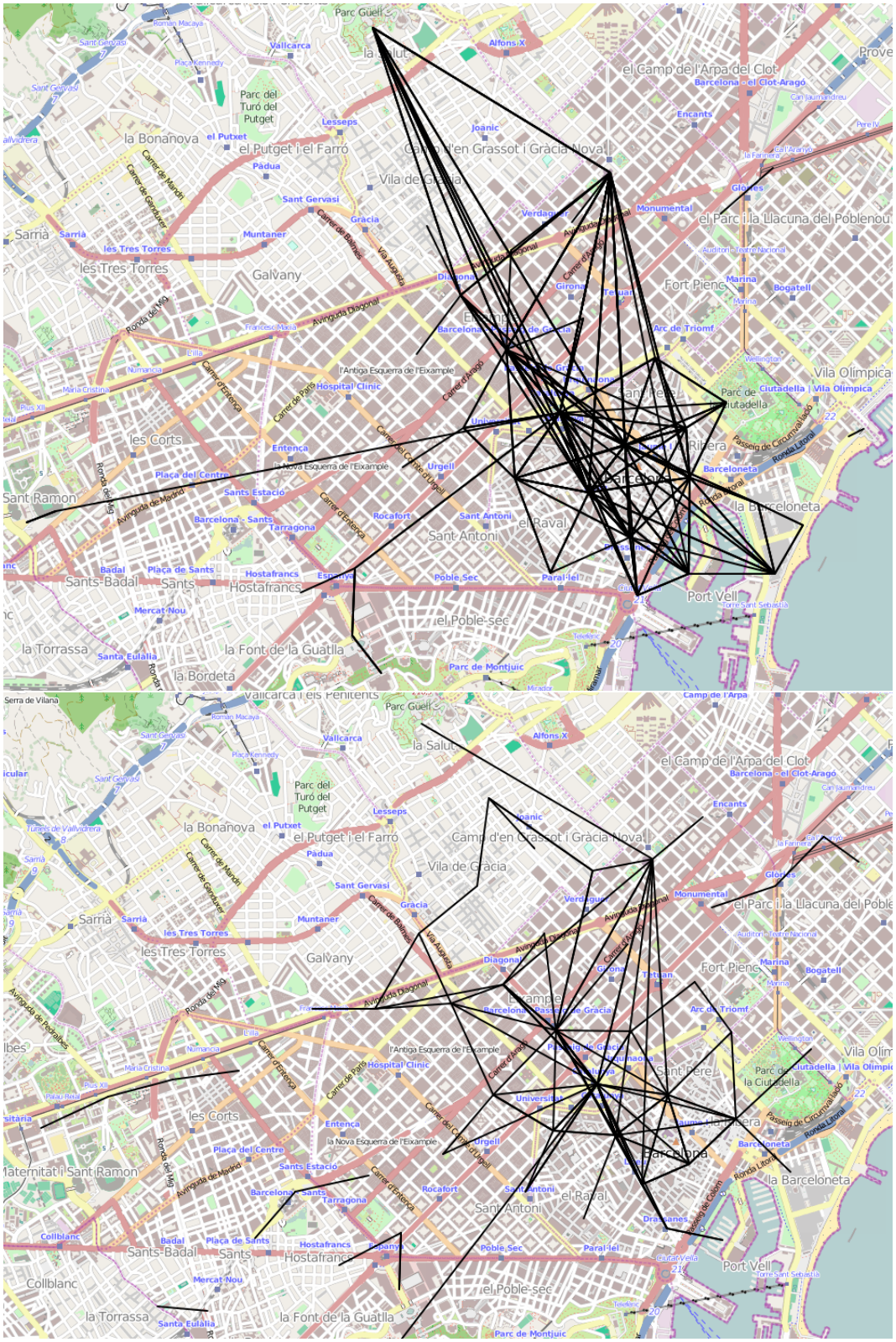

Fig. 9: Frequent connected areas from Barcelona Instragram (up) and Twitter (down) data (Leader Clustering 500 meters, four time intervals, support = 100)

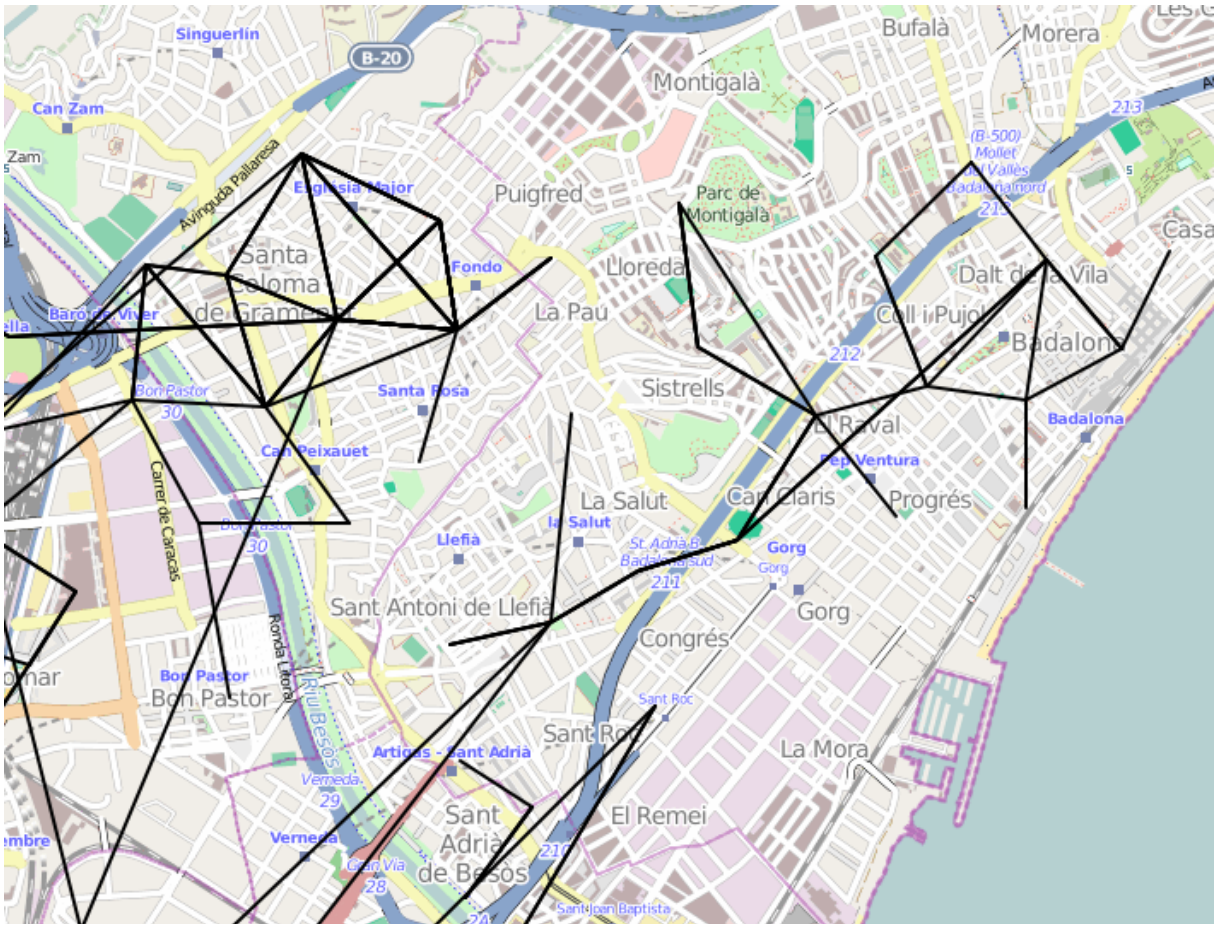

Fig. 10: Frequent connected areas for cities in the metropolitan area of Barcelona (Twitter data, Leader Clustering 500 meters, two time intervals, support = 25)

like parks, some churches, commercial streets, a large hotel in the outskirts of the city and the Milan central train station. Probably this train station is the usual way for most of the tourists of arriving to the city. Twitter data findings are similar than for Barcelona, some connections with cities around Milan, some chained patterns around some of main entrances to the city and also connections from/to the main train stations, specially with the largest one. From the Twitter patterns, it is surprising for a non expert on this city and considering it as a prototype of radial city for its shape and streets distribution, to see that most of them are aligned on a north-east to south-west axis that passes through the center of the city. This could indicate some preference of routes around this axis for the people of the city or perhaps is due to the distribution of business and workplaces inside the city.

For more insight into the generated patterns, actual expert knowledge about the cities is necessary to analyze and decide about the novelty or importance of the patterns that appear while changing the parameters of the algorithms, specially when the number of patterns increases. On the one hand, it is expected that obvious patterns appear during the process, however it is also interesting to quantify their support with real data in order to prioritize the possible actions that can be derived. On the other hand, some lesser known patterns also will appear, like for example the Barcelona Twitter data shows a connection between the areas of the city where usually large business events and conventions are hosted and some areas that concentrate hotels where these people usually stay, like the Fira de Barcelona (large conventions center) and the Diagonal Mar area (Marriott, Princess and Hilton hotels), that surprisingly are at opposite sides of the city.

7 Conclusions and future work

Location Based Social Networks are an important source of knowledge for user behavior analysis. Different treatments of the data and the use of different attributes allow to analyze and study the patterns of users in a geographical area. Methods and tools for helping to analyze this data will be of crucial importance in the success of, for example, smart city technologies.

In this paper we present a methodology able to extract patterns that can help to make decisions in the context of the management of a city from different perspectives, like mobility patterns, touristic interests or preferred city routes. The patterns extracted show that it is possible to obtain behavior information from LBSN data. Increasing the quantity and the quality of the data will improve further the patterns and the information that can be obtained.

As future work, we want to link the information of these different networks to extract more complex patterns. The data from Twitter includes Foursquare check-ins, this allows to tag some of the events to specific venues and their categories, allowing for recommender systems applications and user activity recognition and prediction. There are also links to Instagram photographs allowing to cross reference both networks augmenting the information of user Twitter events with Instagram events of the same user, reducing this way the sparsity of the data. Also, in this paper, the temporal dimension of the dataset has not been fully exploited. Analyzing the events temporal relationship will allow the study of causal dependencies and temporal correlations.

8 Acknowledgments

This work has been supported by the EU funded SUPERHUB project (ICT-FP7-289067).

References

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[2] Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, Dino Pedreschi, and Fosca Giannotti. Interactive visual clustering of large collections of trajectories. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 3–10. IEEE, 2009.

[3] Natalia Andrienko and Gennady Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, pages 1–29, 2012.

[4] R. Dubes and A Jain. *Algorithms for Clustering Data*. PHI Series in Computer Science. Prentice Hall, 1988.

[5] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.

[6] Katayoun Farrahi and Daniel Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27, January 2011.

[7] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[8] Kenneth Joseph, Chun How Tan, and Kathleen M. Carley. Beyond "local", "categories" and "friends": Clustering foursquare users with latent "topics". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 919–926, New York, NY, USA, 2012. ACM.

[9] Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal and ubiquitous computing*, 17(4):605–620, 2013.

[10] Nan Li and Guanling Chen. Analysis of a location-based social network. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, pages 263–270, Aug 2009.

[11] Ross Maciejewski, Stephen Rudolph, Ryan Hafen, Ahmad Abusalah, Mohamed Yakout, Mourad Ouzzani, William S Cleveland, Shaun J Grannis, and David S Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *Visualization and Computer Graphics, IEEE Transactions on*, 16(2):205–220, 2010.

[12] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[13] F. Pianese, Xueli An, F. Kawsar, and H. Ishizuka. Discovering and predicting user routines by differential analysis of social network traces. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a*, pages 1–9, June 2013.

[14] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, pages 1029–1038. ACM, 2010.

[15] Yu Zheng. Location-based social networks: Users. In *Computing with Spatial Trajectories*, pages 243–276. Springer, 2011.