Extracting User Spatio-Temporal Profiles from Location Based Social Networks

Javier Béjar
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya
bejar@cs.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 mobility profiles, points of interest or unusual events. This information is important for different domains like mobility 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 how the spatio-temporal behavior of a large number of users in a well limited geographical area can be segmented in different profiles. These behavioral profiles are obtained by means of clustering algorithms that show the different behaviors that people have when living and visiting a city.

The data analyzed was obtained from the public data feeds of Twitter and Instagram inside the area of the city of Barcelona 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 any general area) to discover useful profiles that can be described on terms of the city singular places and areas and their temporal relationships. These profiles can be used as a basis for making decisions in different application domains, specially those related with mobility inside and outside a city.

Keywords: Spatio-Temporal Data, Clustering, Location Based Social Networks, Smart Cities, User Profiling

1 Introduction

Location Based Social Networks [18], like for example Twitter or Instagram, are an important source of information for studying the geospatial and temporal behavior of a large number of users. The data that these networks provide include the spatio-temporal patterns that users generate while interacting with the different locations inside a geographical area and the events that occur within it. That information can be used to uncover different complex behaviors and patterns, including frequent routes, points of interest, group profiles or unusual events. To study these patterns could be an important source of knowledge for applications such as city management and planning decision support systems or different kinds of recommender systems in route planning and touristic domains.

The goal of this paper is to analyze these spatio-temporal data using different clustering algorithms in order to find out what collective patterns arise from user behavior in large cities. The data used in this analysis was obtained from Twitter and Instagram social networks in the geographical area that surrounds the city of Barcelona. We expect that the results of this study can be generalized to the analysis of data from any city (or general area) so other useful patterns can be discovered from these or similar sources.

The aim is that the spatio-temporal patterns and behaviors discovered could be used later for application domains that need structured information about the behavior of the dwellers of 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 the optimization of 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 with the goal of implementing route planning and mobility recommendation systems.

The plan of the paper is as follows: Section 2 introduces to other approaches to discover patterns/profiles 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 clustering algorithms, to discover clusters as an approximation to the behavioral profiles of the users and its relation to the points and regions of interest in the city. Section 5 shows the results obtained of applying the described techniques to the data collected for Barcelona from Twitter and Instagram. Finally in section 6 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 ranging from GPS traces extracted from these devices to 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 an specific domain of analysis, for example, city officials studying citizen mobility and traffic distribution. In this line of work, [1, 2] 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 [6] to predict user routines and their social connections using hidden markov and gaussian mixtures models. The same dataset was also used in [7] 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 [17], data collected from GPS trajectories over an extended period of time inside a city was used. From these trajectories, a set of special points named staying points were extracted. These were defined as points inside a bounded region where a user stays for a short period of time. These were the points of interest of the user. The points and the categories of the places inside the region surrounding these points of interest were used as the base for a touristic recommender system.

The main issue with GPS data is that they are difficult to obtain continuously for a large number of users. Also GPS traces are not event oriented, meaning that a large number of points from the trace do not account for relevant user activity, obliging to a preprocess of the traces to identify the relevant events.

An alternative to the information collected by GPS enabled devices are the Location Based Social Networks (LBSN). These social networks allow to sample information from a large number of users simultaneously and in an event oriented way. The user only has a new data point when generates a

new relevant event. The main drawback is that the sampling frequency is much lower (only a few events per day) so certain analysis are more difficult or impossible. Also this data source is sparser, not all the events generated by the user are registered.

There are different works that extract patterns from LSBN data. The previously mentioned text mining analogy is used in [9] 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 [14], 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.

3 The dataset

The aim of this paper is to extract useful clusters from LBSN that could be used for the analysis of the behaviors of people living and visiting a city. 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 twelve months. A priori, the quality of these feeds can be considered as non optimal due to the limitations to availability imposed by these social networks for free access data. 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. The analysis described in [12] shows that the sampling provided by Twitter 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 data collection is large. This period is of a complete year of data 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 common 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 previously mentioned, the data obtained from Twitter and Instagram was geographically constrained. The events were filtered to extract only the ones inside an area of approximately 30×30 km^2 of the city of Barcelona. The size of the area was chosen to include all the populated areas of the city and other surrounding cities. This means that also the behavior of the citizens in these other cities is included, allowing to extract not only internal behavioral patterns but also outside behavior and interactions among different cities.

The dataset was collected during a twelve months period (october 2013 to september 2014), the number of events extracted from Twitter and Instagram is around three millions each. 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

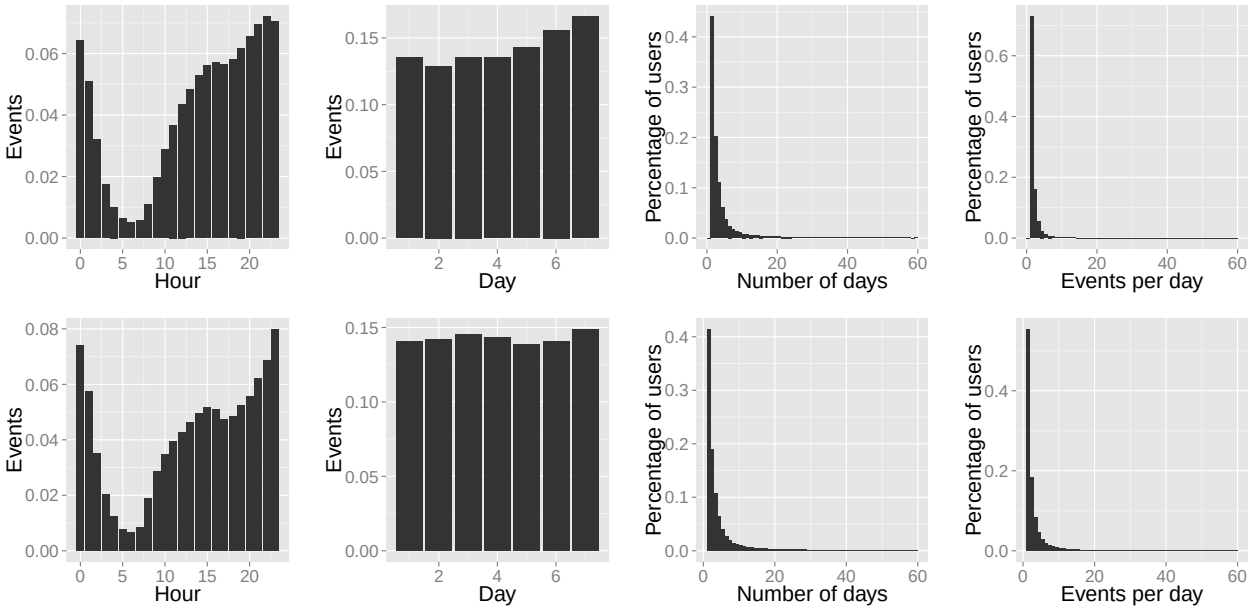

Fig. 1: Hourly and dayweek events, distribution of number of days per user and distribution of the number of events per day (Up: Barcelona - Instagram / Down: Barcelona - Twitter)

a more detailed statistical analysis of the dataset is performed.

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.

Figure 1 shows different information of the events from the datasets collected for the city of Barcelona. The first plot of the figures represents the hourly percentage of events. Both social networks have a similar distribution, the distributions show two separated modalities, one centered around 2-3pm and another centered around 10pm. This means that there are two distinct 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 4-5pm that marks the beginning and the end of these behaviors. A reasonable explanation is the daily work-leisure cycle. During work hours there is less people with the time to generate events and during leisure time, people 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 is 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 a small text than to take a photograph.

The third plot is the percentage of users respect to the number of days when 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 at 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 of the statistical distribution. The actual distribution with all the data 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 during a day, being around 50% for Twitter. The reason could also be that it is easier and faster to write a text than 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 during a day.

4 Clustering user events

The discovery goal is to obtain groups of users that show similar behavior and that can be interpreted as user profiles. To use the events of individual user days as dataset would result in very simplistic patterns, given the sparsity of the data. A user day usually contains less than three or four events. To obtain more complex patterns, it was decided that the behavior of a user during all the study period would be more informative. The history of events of a user would represent better his individual behavior profile. Also, summarizing this way the user events will result in very different examples when the events come from users that visit the city for a short period of time or from users that actually live in the city. This will help to separate more clearly these very different kinds of users and will provide with clusters more easy to interpret and classify.

Before the clustering can be performed, some transformations and decisions have to be taken concerning to what attributes will be used for describing the users, what values will be used to represent them and what clustering algorithms to use for obtaining the user profiles. All this problems will be addressed in the following subsections.

4.1 Data preprocess

As mentioned previously, it was decided to summarize all the events for the users collected during the all data gathering period, summing up all the different places where they have been and when. It is difficult to extract patterns of the behavior of the users directly from the raw events. Given that the geographical positions correspond to point coordinates inside the area, the probability of having a large number of events at the same coordinates for several users is extremely low. The same problem appears for the temporal dimension. This means that the geographical positions and time dimension have to be discretized in some way to increase the similarity and the probability of coincidence of the events. This will make the attributes to represent the occurrence of an event inside a geographical area during an specific range of hours of the day.

Different discretizations of a geographical area can be proposed to allow the extraction of patterns at different resolutions and complexity. A simple approach that has been used in previous analysis of this dataset (see [4]) is to divide the area using 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, reducing the total number of possible attributes.

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 the counts for some events could be split into adjacent cells by the discretization, this makes this method less reliable than other alternatives that can obtain a discretization that adapts better to the actual densities that appear in the data. A larger granularity could perhaps reduce this problem but with the cost of the loss of information.

An alternative method is to use clustering algorithms for the discretization. These algorithms can approximate better the different geographical densities of the events. From all the clustering algorithms that can be applied, a simple and interesting possibility is to use an incremental clustering algorithm, like the leader clustering algorithm [5], 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.

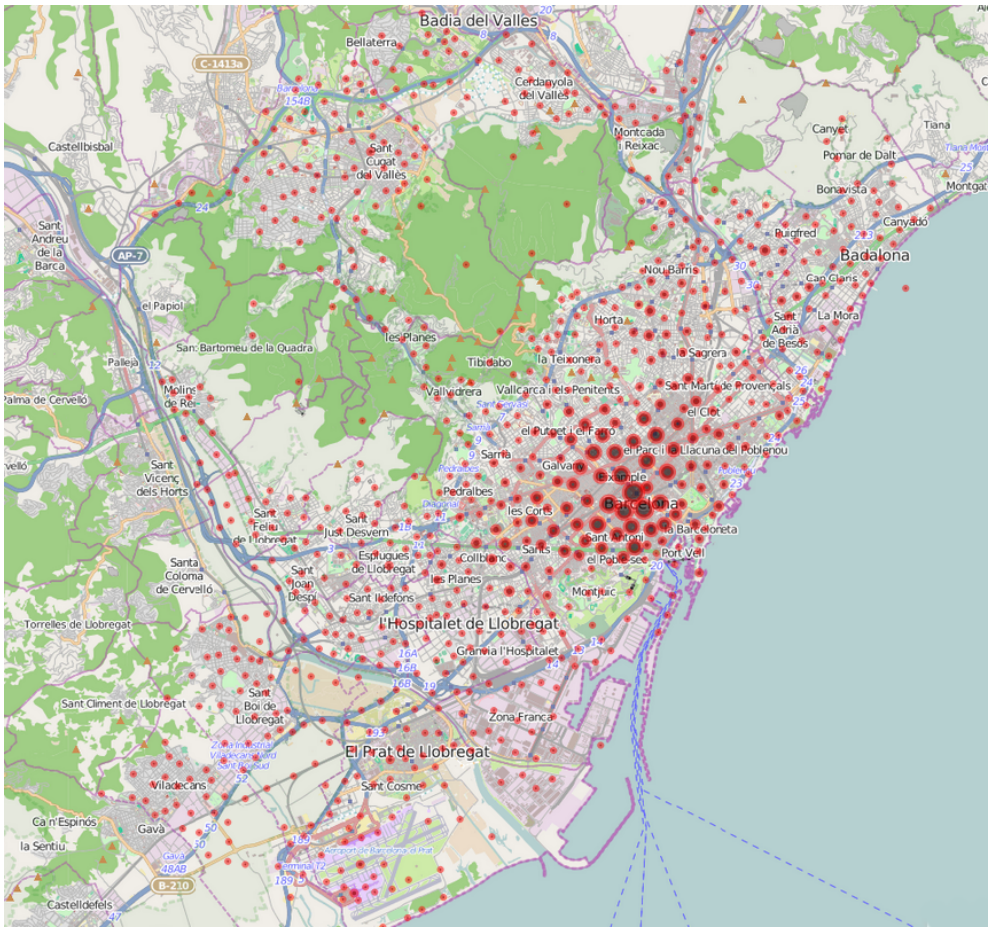

Fig. 2: Clustering of Barcelona Twitter events (radius=500m, more than 25 tweets) the size of the clusters is proportional to the number of events in each cluster

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 consist in that the clusters adapt to the different densities of the data and that it also can be computed in linear time respect to the number of events. Figure 2 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 into several clusters. Applying this algorithm the spatial coordinates of the events can be grouped and the centroid of each cluster can be used as representative.

Another possible clustering algorithms to apply would be 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 is a very large dataset, it could arise scalability issues. Another practical problem is to tune the parameters of these algorithms for obtaining a satisfactory discretization. The particularity of this data makes a vast majority of the events to be concentrated in specific areas and depending on the parameters of the clustering, very large clusters appear covering very extensive areas and concentrating most of the events. This kind of dicretization would result in trivial clusters.

For the experiments and following the results presented in [4] where this discretization methods were applied for the same domain, but for obtaining different kind of patterns (frequent itemsets), it was decided to use a discretization based on the leader clustering algorithm.

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

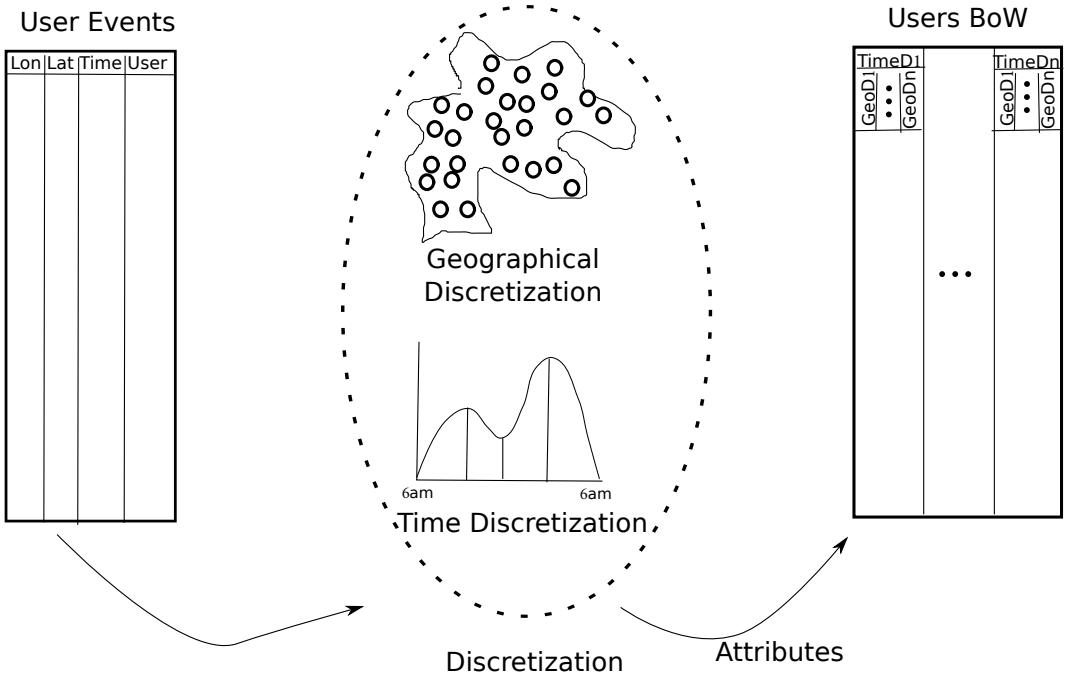

Fig. 3: Summarizing user events using a bag of word representation from the data discretization

discretization meaningful for the data. As we have shown in 3.1, there are two modalities in this distribution with means around 4pm and 10pm 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 from the events perspective a day begins and ends around 6am. This means that a daily transaction has to include events within this interval of hours to be correct.

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.

Using these transformations we can build a dataset whose attributes are defined by the possible geographical areas (defined by the geographical discretization method and the discretization granularity) and by their timestamp (defined by the time discretization method and the time intervals). Generating specific values for these attributes a dataset can be obtained to which different clustering algorithms can be applied to uncover the behavioral profiles according to the user similarity.

4.2 Dataset representation and attribute values

The total number of possible attributes will vary with the choice of discretization granularity but it can range from a few thousands for very coarse granularity to several tens of thousands for more fine granularity. Given that the total number of events that a user has is a small number respect of the total number of attributes, we will have a very sparse dataset. In order to choose an adequate representation for this dataset it was decided to use the text mining analogy already used on related

work (see for example [7]). In our case we can make the analogy of user events with words and the collected behavior of each user as documents.

To obtain the summary for each user behavior, a feature vector is generated using the vector space model/bag of words (BoW) following the geographical and time discretizations (see figure 3). For the attribute values, we have to compute the term frequency (TF) and inverse document frequency (IDF) [16]. There are different term/event frequency values that can be used. Being the task at hand exploratory, three different possibilities widely used in text mining have been evaluated:

1. Absolute term frequency, computed as the times the user has been in a area during an specific time interval.
2. Normalized term frequency, computed as the times the user has been in a place during an specific time interval, normalized by the total number of areas the user has been.
3. Binary term frequency, computed as 0 or 1, depending on whether the user has been or not in a certain area during an specific time interval.

To include in the representation the importance of the places on the city respect to the global number of visits they have, also the inverse document frequency (IDF) was computed for all the different places/times in the dataset. This allows to obtain six different representation of the users data, one for each type of term frequency attribute and one for each respective IDF normalization.

4.3 Clustering algorithms

Different cluster algorithms can be applied to extract group profiles. Due to the representation of the data (bag of words), our intuition is that clustering algorithms usually applied for this representation would be successful in finding meaningful clusters. To test this intuition, we experimented with three different clustering algorithms, K-means [3], spectral clustering [13] and affinity propagation clustering [8].

K-means is based on finding spherical clusters around a prototype. It has an acceptable computational complexity being able to work with sparse data as is our case. The main issue for this method is how to decide the correct number of clusters. This task is harder because the assumption of spherical clusters is probably incorrect for most of the clusters in the data, so the usual quality indices employed to decide the number of clusters will mot be very useful. This arises the need for experimenting with different numbers of clusters and to evaluate other subjective characteristics of the clusters. Also the clusters will be probably difficult to separate because some users will have a behavior that is a mixture of different behaviors, K-means obtains a hard clustering of the data. An adequate approach would be to obtain a large number of clusters that could be grouped after using other clustering algorithm.

Affinity propagation is an exemplar based clustering algorithm based on belief propagation. In this case the beliefs are related to the ability of an example to represent the examples that are close (availability) and the belief of the example that a particular example represents them well (responsibility). The algorithm uses message passing that updates these beliefs until convergence and the initial beliefs are obtained from the examples similarities. This algorithm is also able to work with sparse data, and has been successfully used for text mining tasks, but its computational complexity can be almost quadratic. Its main advantages respect the first alternative is that it is able to find irregular shaped clusters and also that is able to decide the number of clusters that best fits the data.

Spectral clustering is a graph based clustering algorithm that uses the graph Laplace matrix computed from the similarity matrix of the examples. The eigenvectors of the Laplace matrix are obtained and used as a transformation from the original data. This transformation maintains the local structure of the data allowing to discover non spherical clusters in the original dataset. After the transformation different algorithms can be used to obtain the clusters, for example K-means. The computational cost of the algorithm depends on the eigendecomposition but the graph spectral matrix is in our case very sparse (few examples are similar enough to each other), so the cost is in the same complexity as affinity propagation. Also this algorithm has been applied successfully for datasets with a bag of word representation.

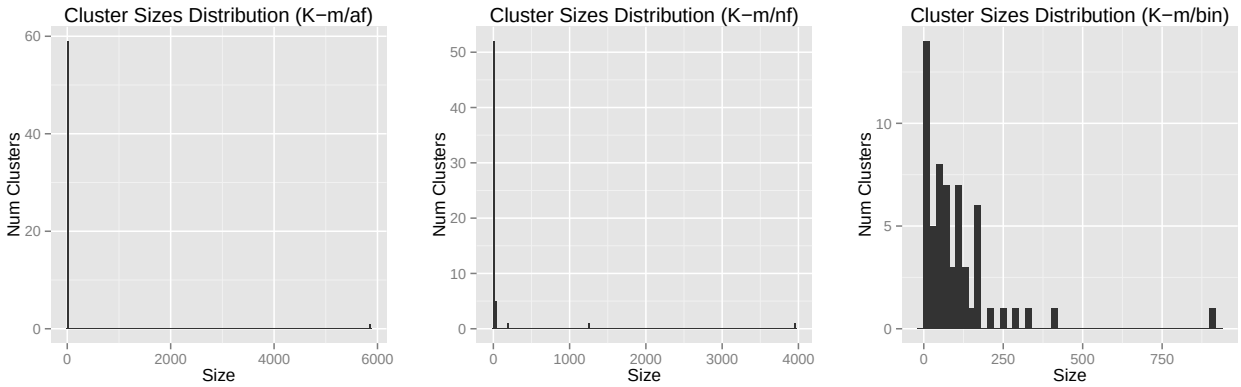

Fig. 4: Cluster sizes distribution using Twitter data (discretization: two time intervals, 250m diameter) for k-means with 60 clusters using absolute frequency, normalized frequency and binary frequency attributes)

5 Experiments

In order to enhance the quality of the dataset, we filtered users without a minimum number of distinct events (place/time). This allows to extract more meaningful profiles with the cost of reducing the actual number of users. Given that the dataset is sparse, a large portion of users has not been captured a significant number of times during the collection period, so makes sense to discard this information for our purposes. Also, given that the collection period (a year) is long, there is a large confidence that the behavior of users with more than a threshold of different events have been captured.

In the experiments, we have used a threshold of at least 20 different events, considering as different being inside the same region but at a different time slot. This value reduces the number of users depending on the space and time granularity and the dataset to around ten thousand users in both datasets. We consider that this number of users is significant enough to show very different profiles.

In the clustering results, certain number of small sized clusters are bound to appear for profiles not very represented in the data. As a quality criteria we have considered that a cluster is significant if has a minimum user support (at least 20 users in our experiments), discarding the clusters with less users as noise.

To evaluate the quality of the clusterings we have considered two subjective criteria. The first one is that the clustering has to result in a large number of clusters. Given that we are grouping several thousands of users it is more reasonable to assume the existence of many different behaviors. The second one is that the distribution of the sizes of the clusters has to include large clusters for more common behavior (tourists, for instance) but also small and medium sized clusters for more specific behaviors.

5.1 The attribute values

As previously mentioned, we have chosen three possible different term frequency values for the attributes in the bag of words representation with corresponding IDF normalization. The experiments with the different types of values for all the datasets show that the absolute term frequency and the normalized term frequency (with and without IDF normalization) do not result in a good representation of group behavior given the small number of clusters with a size larger than the support.

In figure 4 is represented the histogram of the sizes of the clusters using K-means for obtaining 60 clusters using Twitter data with a specific space and time discretization. Using the absolute term frequency (the number of times an event occurs for a user) as attribute only a cluster with most of the examples is obtained and the rest correspond to clusters with less examples than the selected support. For the normalized term frequency (the number of times an event occurs for a user divided by the sum of the count of his events) only four clusters appear above the support with a cluster with more

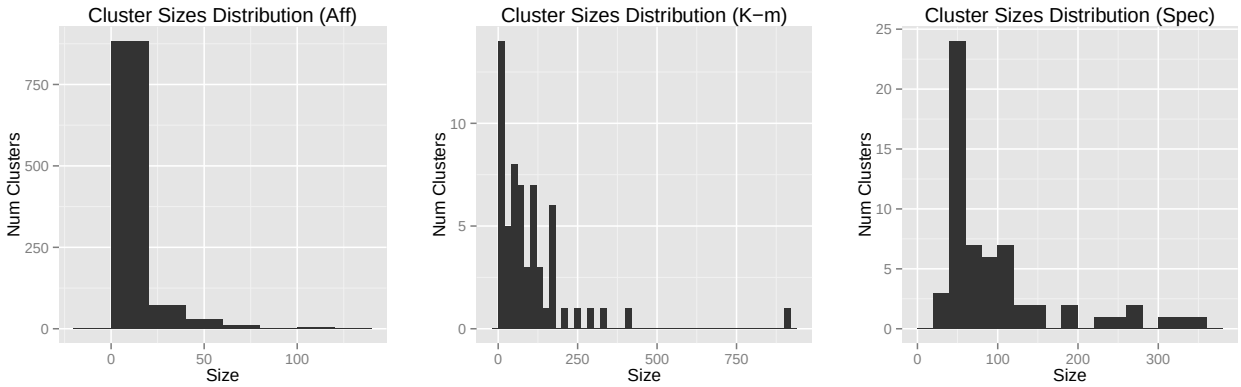

Fig. 5: Cluster sizes distribution using Twitter data (discretization: two time intervals, 250m diameter) for k-means and spectral clustering with 60 clusters and affinity propagation with damping factor 0.5 using binary frequency attributes)

than two thirds of the examples. With the binary term frequency (the event has occurred or not for the user) there are more than forty clusters with a wide range of users for this discretization. Similar results are obtained with different data discretizations and clustering algorithms with this dataset and also with the Instagram dataset.

Given these results only the binary term representation will be considered for further experimentation.

5.2 The clustering algorithms

Given the different assumptions and bias of clustering algorithms, an analysis of its different results respect to the kind of clusters obtained is needed. In this section, the distribution of the sizes of the clusters and the similarity of the clusterings will be analyzed.

Affinity clustering automatically determines the adequate number of clusters for the data. This only depends on one parameter of the algorithm, the damping factor, that controls how much the different messages that are used to decide the assignment of the examples are updated each iteration. With our datasets, this algorithm always returns a very large number of clusters, depending on the discretization of the data, that ranges from around 600 to 1100 clusters. A large proportion of this clusters only have one instance or are below the support threshold, leaving with around 75 and 100 not very large clusters that represent specific behaviors. It was expected to find a large number of clusters given that it is more plausible that several thousands of users picked at random will show a large variety of group behaviors. Although, larger groups were expected for more common behavior.

K-means needs the number of classes to be specified. For the experiments and given the number of clusters over the support obtained by affinity clustering the range between 60 and 100 was considered to obtain the expected clusters. The results from the experiments show that a large portion of the clusters only contain one example or are under the support as happens for affinity clustering. The final number of clusters depends largely on the space discretization. For the range of target number of clusters, with a discretization of 100m are discovered between 30 and 40 clusters, for 250m are discovered between 45 and 55 clusters and for 500m are discovered between 50 and 75 clusters. The distribution of sizes is more reasonable having a very large cluster that includes most of the tourists in the dataset. Also a variety of specific and general clusters appears. Usually when a larger number of clusters is used, some of the small clusters are split, remaining the larger clusters intact.

The implementation used for spectral clustering also uses K-means as the clustering algorithm for the post process of the dataset after transformation using the Laplacian matrix. The same range for the number of clusters than for K-means was used. The results from the experiments show that there are almost no clusters under the support threshold and the sizes of the clusters decreases with the target number of clusters, so large clusters are split when more clusters are demanded. In this case the

(2T/250m)		Affinity		K-means			Spectral		
	AMI	0.5	1	60	80	100	60	80	100
Affinity	0.5	-	0.42	0.15	0.17	0.17	0.22	0.22	0.22
	1		-	0.17	0.18	0.17	0.22	0.22	0.22
K-means	60			-	0.41	0.40	0.31	0.28	0.26
	80				-	0.39	0.32	0.29	0.28
	100					-	0.33	0.30	0.28
Spectral	60						-	0.60	0.53
	80							-	0.68
	100								-

Tab. 1: Cluster similarity among different clustering algorithms using AMI index for Barcelona Twitter data with two time interval discretization and 250m space discretizations

distribution of the sizes of the clusters is more homogeneous and there is a tendency towards smaller clusters, with almost half of the clusters with a size below 100 users, tendency that increases with the target number of clusters.

Figure 5 shows the distribution of the sizes of the clusters for Barcelona Twitter data with a discretization of 250 meters and two time intervals. It can be seen the different distribution of cluster sizes, that evidences that each algorithm extracts a different view of the profiles of the users.

In order to measure how much is shared among the clusters obtained with the different clustering algorithm, external validity measures are a useful tool. In this case, it was decided to use three measures commonly used in the cluster literature, namely Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI) (see [15]). Basically these measures compute the coincidence of pairs of assignments between two partitions. They can be used to compare with a reference partition or as a relative measure among different partitions. The main difference among the three measures is that the last two are adjusted for chance, so the effect of randomness is discounted. All three measures are in the range [0,1], indicating a value of 1 identical partitions.

From the results, time discretization does not seem to affect much to the similarity among the clusterings. The space discretization increases the similarity among the cluster when is coarser. Table 1 shows the values for AMI measures using two time interval discretization and space discretizations of 250m. From the value of the measures, it looks that the similarity among the clusterings is not large, probably due to the large number of clusters, specially for affinity propagation. There is no agreement among the three measures about the similarity among the partitions from the three algorithms, the AMI measure indicates that K-means and spectral clustering results are more similar to each other and equidistant to affinity clustering, ARAND considers K-means and spectral clustering more different to each other and equidistant to spectral clustering, and NMI consider the three algorithms almost equidistant.

The conclusion is that each algorithm obtains a different view of the datasets, needing a visual exploration of the clusters by part of an expert in the domain for the validation of the results.

5.3 Clusters interpretation

In order to facilitate the interpretation of the clusters by the expert, a prototype, represented over a map, is computed as the absolute frequency of the visits of the users to the different clusters of the discretization. This representation can be obtained without considering the time slot of the day to be able to see what places are more visited by the users of a cluster. Also, for a more complex interpretation, a representation that includes the absolute frequency of visits break down by time discretization can be computed. This representation allows to see geographical behavior associated with the time of the day.

Inspecting visually all the clusters obtained using the three algorithm, despite the lower similarity

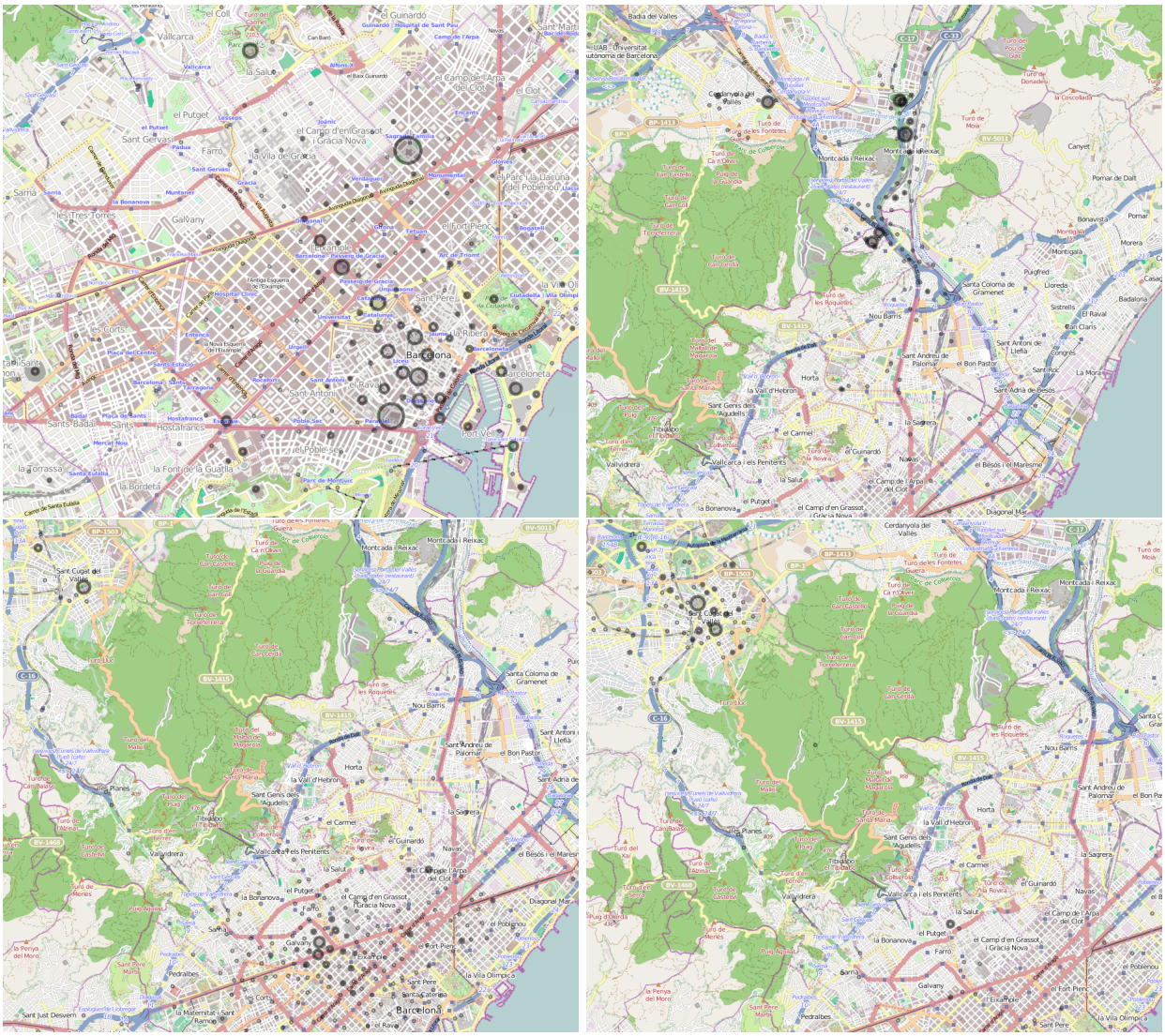

Fig. 6: Clusters obtained by K-means for Barcelona Twitter data showing a large cluster (up, left) described mainly by touristic points of interest, a cluster that shows a rush hour pattern where the more frequent places are concentrated along two highways that enter Barcelona from the north-west (up, right), a small cluster that shows a mobility pattern of people that moves from their home in a Barcelona nearby city to Barcelona (down, left) and a similar clustering for the same nearby city where there is no outside mobility (down, right).

indicated by the cluster validity indices, a lot of common clusters appear. They can be identified because they share the same high probable places or are contained inside similar geographical areas, presenting only differences in the small probability areas that describe them. Despite of that there are also clusters that make sense on the eyes of the experts that appear only for a particular clustering algorithms.

Also using a different number of clusters allows to look to the profiles at different levels of granularity. For this purpose spectral clustering shows a better performance, because it usually splits larger clusters when a larger number of clusters is pursued, allowing to look for more specific profiles.

It is difficult to interpret clustering results without a more profound knowledge of the domain, but from the visualization of the prototypes some evident clusters appear that can be classified in four types. First, clusters with popular behavior with a large number of users, for instance, different clusters that include different subsets of touristic points of interest are recovered by all three clustering algorithms. Second, geographically localized clusters, medium sized clusters that include people that live in an specific suburb of the city or a surrounding city. These users generate events around where they live, usually during leisure hours. Third, geographically dispersed clusters, smaller clusters that show large frequency events at different and distant places all over the studied area, usually associated with mobility patterns inside and outside the city where some of the places with larger frequency are close to public transportation stops or follow specific roads. Four, event specific behavior clusters, small clusters with one or few frequent events and a large number of low frequent events dispersed around a large area, like people arriving or departing from the airport or rush hour events. Figure 6 shows some examples of these kinds of clusters.

6 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 preferred mobility patterns, event profiling, gathering patterns or touristic interests. 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.

7 Acknowledgments

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

References

[1] 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.

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

[3] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *SODA*, pages 1027–1035. SIAM, 2007.

[4] Javier Bejar. Mining frequent spatio-temporal patterns from location based social networks. Technical report, Technical University of Catalonia, LSI-14-10-R, October 2014.

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

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

[7] 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.

[8] Frey and Dueck. Clustering by passing messages between data points. *SCIENCE: Science*, 315, 2007.

[9] 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.

[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] A.Y. Ng, M.I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[14] 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.

[15] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 9999:2837–2854, December 2010.

[16] Sholom M Weiss, Nitin Indurkhya, and Tong Zhang. *Fundamentals of predictive text mining*, volume 41. Springer, 2010.

[17] 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.

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