To appear in the *Journal of Experimental & Theoretical Artificial Intelligence*
Vol. 00, No. 00, Month 20XX, 1–16

Discovery of Spatio-Temporal Patterns from Location Based Social Networks

J. Béjar, S. Álvarez, D. García, I. Gómez, L. Oliva, A. Tejeda, J. Vázquez-Salceda

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Location Based Social Networks (LBSN) like Twitter or Instagram are a good source for user spatio-temporal behavior. These networks collect data from users in such a way that they can be seen as a set of collective and distributed sensors of a geographical area. A low rate sampling of user's location information can be obtained during large intervals of time that can be used to discover complex patterns, including mobility profiles, points of interest or unusual events. These patterns can be used as the elements of a knowledge base for different applications in different domains like mobility route planning, touristic recommendation systems or city planning.

The aim of this paper is twofold, first to analyze the frequent spatio-temporal patterns that users share when living and visiting a city. This behavior is studied by means of frequent itemsets algorithms in order to establish some associations among visits that can be interpreted as interesting routes or spatio-temporal connections. Second, to analyze how the spatio-temporal behavior of a large number of users can be segmented in different profiles. These behavioral profiles are obtained by means of clustering algorithms that show the different patterns of behavior of visitors and citizens.

The data analyzed was obtained from the public data feeds of Twitter and Instagram within an area surrounding 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 singular places and areas and their temporal relationships.

Keywords: Spatio Temporal Data, User Profiling, Data Mining, Frequent Itemsets, Clustering, Location Based Social Networks

1. Introduction

Location Based Social Networks (Zheng, 2011), 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 they 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 for route planning and touristic domains.

The goal of this paper is to analyze these spatio-temporal data using frequent itemsets and clustering algorithms in order to find out what 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 cities of Barcelona and Milan.

Corresponding Author: J. Béjar, Computer Science Department; Universitat Politècnica de Catalunya; E-mail: bejar@cs.upc.edu.

The expectation is that these unsupervised techniques could be applied to any city (or general area) in order to discover useful patterns.

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 the 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 from spatio-temporal data in general and from LBSN in particular. Section 3 describes the characteristics of the data and the transformations applied to obtain datasets suitable for the discovery goals. 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. Section 5 shows the results obtained of applying this technique to the data from the different datasets. Section 6 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 7 shows the results obtained from applying these algorithms to the same data sets. Finally, section 8 presents some conclusions about the results of the different techniques 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. 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 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, (Andrienko & Andrienko, 2012) describe different methods for obtaining visualizations of clusters of GPS trajectories extracted from the movements of cars inside the city of Milan.

Other applications include user routine mining and prediction. The idea is either to recognize user activities from repeating temporal behavior, or to recommend activities according to past behavior. Data from mobile phones of MIT students and faculty was used in (Eagle & Pentland, 2006) to predict user routines and social connections using HMM and gaussian mixtures models. The same data was used in (Farrahi & Gatica, 2011) for user and group profiling, routine prediction and change 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 activities to a bag of words representation and to cluster them using Latent Dirichlet Allocation.

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

the relevant events. An alternative the Location Based Social Networks (LBSN). These 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 so certain analysis are more difficult. Also this data source is sparser, not all the generated events are registered.

There are different works that extract patterns from LSBN data. The same text mining analogy is used in (Joseph, Tan, & Carley, 2012) to analyze data from Foursquare. 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 characterized the different groups of persons in a city, as a first step to extract user profiles. In (Pianese, An, Kawsar, & Ishizuka, 2013), data from Twitter cointaining Foursquare check-ins was used to predict user activity. 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. In (Lee, Wakamiya, & Sumiya, 2013) 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 each 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 behavioral patterns in each cluster. As an a posteriori analysis, the categories and distribution of the venues in the clusters, given the frequent patterns, were used for their characterization.

3. The dataset

The aim of this paper is to extract patterns from LBSN useful for the analysis of the behavior of people living and visiting a city. Our interest is focused on data 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. All the events obtained (tweets, photographs) include spatio-temporal information represented as latitude, longitude and timestamp. A unique user identifier is also provided for each event that allows to relate all the events of a user. This information represents a low rate sampling of the spatio-temporal behavior of a large number of users.

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 of a maximum of the 1% of the tweets at a given moment of time. There has been some studies of the quality of this specific data source. The analysis described in (Morstatter, Pfeffer, Liu, & Carley, 2013) 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 with a longer period of data collection. Due that we are interested in the common activity of users, and that the period of data collection was at least six months long, we consider that it is representative enough to provide meaningful results.

The data obtained from Twitter and Instagram was geographically constrained. The events were filtered to the ones inside an area of 30×30 km^2 around both cities. The size of the area was chosen to include all the populated areas of the cities 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

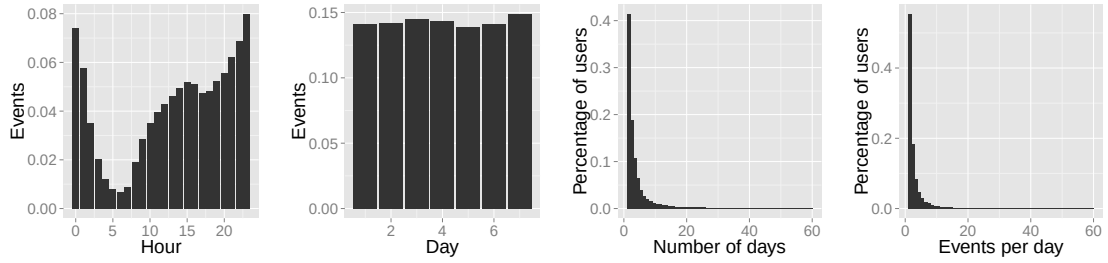

Figure 1. Hourly and dayweek events, distribution of number of days per user and distribution of the number of events per day for Barcelona Twitter dataset)

behavior and their interactions.

The dataset for Barcelona 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. 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 visits Barcelona.

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.

3.1. *Statistical analysis*

In order to understand the characteristics of the data, an statistical analysis was performed, including the frequency of events for natural periods like weeks and days. 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 were analyzed.

Figure 1 shows an example of the results obtained from the analysis of the events for the Barcelona Twitter dataset. The first plot represents the hourly percentage of events. All datasets have a similar distributions, there are two separated modalities, centered around 2-3pm and 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 plots, showing a decrease of activity at 4-5pm that marks the beginning and the end of these behaviors. An 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. Instagram plots show that this network is more used during the weekends compared to Twitter. An explanation is that, being a photo sharing social network, it is more probable to have something to show during weekends than during working days. This trend 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 shows the percentage of users with respect to the number of days when they have any activity during the collected period. Given that the data are a random sample of the actual events, the probability of capturing repeatedly a casual user is very small. Also, the tourists visiting these cities make 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. For both networks, the distribution of the percentage decreases with the number of days following a power law.

The fourth plot is the distribution of how many events usually a user generates in a

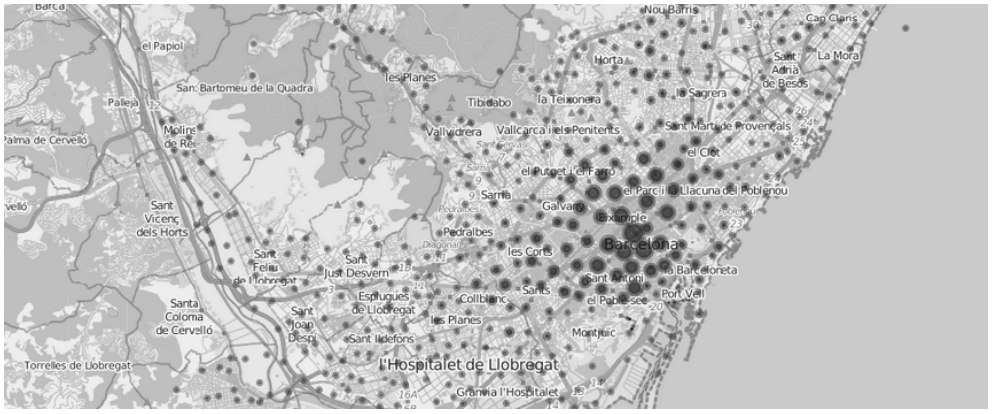

Figure 2. Clustering of Barcelona Twitter events (diameter=500m, more than 25 tweets), the size of the clusters is proportional to the number of events in each cluster

given day. Because only a random subset of the events is captured, most of the users will have a number of events very low. 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.

3.2. *Data preprocessing*

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 areas, the probability of having a large number of events at the same coordinates for several users is extremely low. The problem is worse if the temporal dimension is included. 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 of the data to represent the occurrence of an event inside a geographical area during an specific range of hours of the day.

Different geographical discretizations can be proposed to allow the extraction of patterns at different resolutions. A possibility is to divide the area using a regular grid. This would also allow 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 to organizationally meaningful regions. The main drawback is that some associations could be lost depending on how the events are divided into adjacent cells.

Given that a regular grid cannot adapt to the different geographical densities of the events, to use a clustering algorithm to generate a discretization that discovers these densities could be an alternative. 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.

It was decided to use the leader clustering algorithm (Dubes & Jain, 1988) 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 positions of the clusters adapt to the densities of the data and that it can be computed in linear time respect to the number of events. Figure 2 shows a clustering 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 events and reducing the possibility of splitting close events in several clusters.

The discretization of time is a more simple issue. To use the hourly distribution of the events seems a reasonable choice for finding a discretization. As shown in 3.1, there are two modalities in this distribution that allow to split the day 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 pattern has to include events within this hours to be correct.

Using these discretizations we can build different datasets. The attributes are defined by the geographical areas (defined by the geographical discretization granularity) and their timestamp (defined by the time intervals). To these datasets we can apply frequent itemsets algorithms to uncover frequent daily patterns. Also, generating specific values for these attributes, datasets can be obtained to which different clustering algorithms can be applied to uncover the behavioral profiles according to the users similarity.

4. Frequent routes discovery

The first goal for the analysis of these datasets is to discover connections between geographical places so the relationships among the events the users generate are revealed. We will consider the individual events as the building blocks of the patterns and the connections among the events as the patterns. The assumption is that if the same set of events, considering an event as being in an area around a 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 associations among places that are visited frequently by people that have a certain profile (e.g. tourists), places in the city that need to be connected by public transportation or traffic bottlenecks that are connected in time.

This discovery goal is similar to the one solved by market basket analysis. In this type of analysis the transactions from the purchases of users during their visit to a store are collected. The patterns discovered are the associations among products defined by the frequent purchase of groups of clients. Analogously, we can define a transaction in our domain 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.

Using the data preprocessing explained in the previous section (3.2), the next subsection will explain how to define the transactions from the LBSN data and the methods used for the pattern extraction.

4.1. *Generating frequent routes*

In our formulation, a transaction contains the daily events (items) for the users at different geographical areas at different ranges of time. It can be considered significant to observe a number of simultaneous events higher than a specific support threshold. These 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 these associations as routes or connections between different geographical points at different points of time.

The main issue for obtaining these associations 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 instance, using a discretization with a diameter of $500m$ usually some thousands of geographical areas are obtained, multiplied by the size of the time interval discretization.

The extraction of these sets of frequent events can be obtained by the application of frequent itemsets algorithms. 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 only can be frequent if all its subsets that contain one element less are also frequent. There are different algorithms for frequent itemsets discovery, for our case, the best approach is the FP-Growth algorithm (Han, Pei, Yin, & Mao, 2004). This algorithm avoids to generate all possible candidates by summarizing the transactions of a database using a specialized data structure called FP-Tree. This data structure uses a prefix tree approach to store the transactions. 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 patterns uses this structure to perform the exploration. Beginning with all items that appear on the first level of the FP-Tree, generates all possible frequent itemsets with the items that 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 an itemset, the FP-Tree is divided and explored in parallel if necessary. This circumstance makes this algorithm very scalable, being able to process datasets with large numbers of items and transactions.

With these algorithms, different kinds of subsets of the total set of frequent itemsets can be extracted. To obtain all possible itemsets with a support larger than a threshold usually results in a very large number of redundant patterns. In our case, these redundant patterns represent all subroutes of a longer route. For the applications that we are interested in, the most suitable subset of itemsets is the one that only contains the longest possible frequent itemsets that have a support larger than the minimum support. This kind of patterns are called *maximal itemsets*. These will be the ones chosen to summarize all the connections between places and time slots and will represent the more general set of areas that are frequently connected during a day.

One issue for these algorithms is how to select an adequate value for the support. Given the sparsity of the data, the selection of the minimum support is a difficult problem. The domain indicates that to be significant, the number of people that presents a pattern has to be large. Although, given that we have an incomplete picture of the users, because of the random sampling, the actual value has not to be too strict.

It also has to be considered that the density of events is different in different parts of the city. As can be seen in figure 2, that represents the discretization proportionally to the 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. Extracting frequent patterns

In this section we explore the different parameters for our approach and interpret the results obtained. For the granularity of the 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 clusters with a very low number of events.

For the time discretization, the events show two distinct hourly distributions during the day (see figure 1). A discretization following these distributions seems the more reasonable choice, having days that begin and end at 6am. In the experiments we have used the following intervals: two ranges, 6am to 6pm and 6pm to 6am of the next day; three ranges, 6am to 4pm, 4pm to 10pm and 10pm to 6am of the next day (the two distinctive populations of events, but separating the range where the populations are mixed); four

	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

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

ranges that splits the first discretization in two at the mean values of the distributions.

Frequent itemsets algorithms use a threshold as significance measure (minimum support). This value has to be defined experimentally guided by the domain knowledge. 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 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. *Algorithms parameters*

The support threshold determines what patterns are significant. Given that the events are a sample 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 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 for this city. The difference is even larger when the discretization is finer, only explained assuming that the patterns are geographically more concentrated for the Instagram dataset due to the larger number of tourists.

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 the decreasing of patterns with the increase of the support also to follow 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. 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 applications, it could make more sense

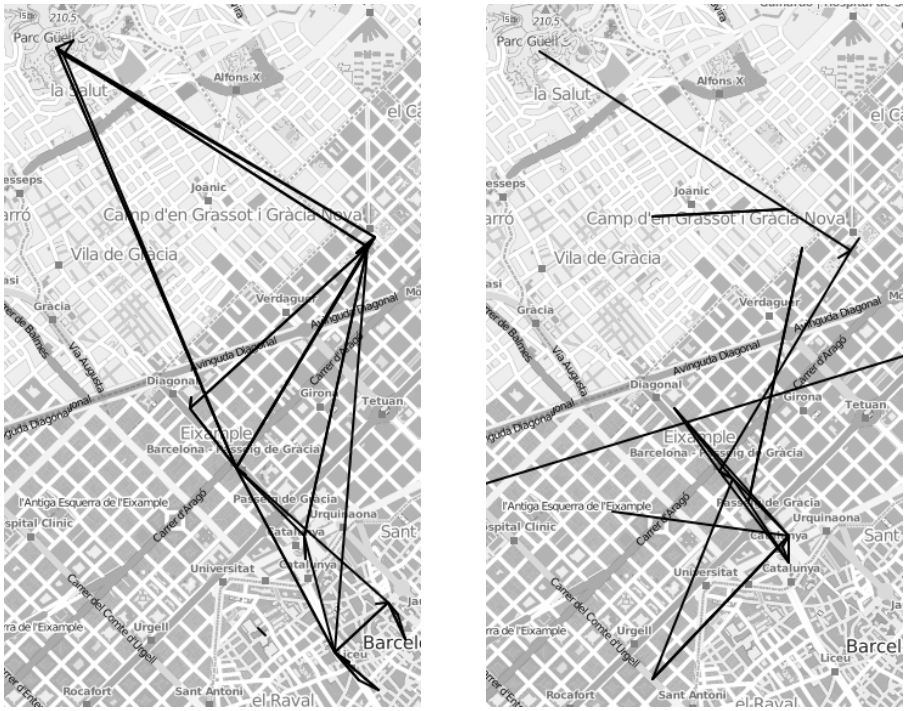

Figure 3. Frequent connected areas from Barcelona Instagram (left) and Twitter (right) data (Leader Clustering 100 meters, two time intervals, support = 100)

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 geographically disperse, 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.

5.2. *Frequent patterns interpretation*

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. 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, 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, showing connections among touristic points of interest. Patterns from Twitter events are more diverse, showing this in the kinds of places that appear connected, that include touristic points of interest but also other types of places distributed all over the city.

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 3, left) that represent the most visited places in the city by tourists and how they are connected. All routes have length two, this is expected because the support is high. Some routes appear several times with different time relations. Twitter for the same parameters reduces the number of routes to 13, it shows tourist activity (figure 3, right), but also other connections appear related with areas of high nightlife activity. In this set of routes there are four with length three, some connect nearby places, and others correspond to return routes.

For these parameter the Milan Instagram generates just two patterns around the Duomo cathedral in the center of the city. Milan Twitter data generates 13 patterns, these are longer than 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 streets around 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. 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 that chains of patterns appear at the main entrances of the city and around the main train station, suggesting rush hour patterns. It 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 only this 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. Connections from the airport also increase, mainly including connections with subway, train and bus stations.

For Milan, Instagram data with a support of 25 events includes also other touristic related areas and the Milan central train station. Probably the usual way for most of the tourists and people from nearby areas of arriving to the city. Twitter data findings have similar characteristics than for Barcelona.

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.

6. Clustering user events

The second discovery goal is to obtain groups of users that show similar behavior, interpreted as user profiles. To use the events of individual user days as dataset, as in the previous goal, would result in simplistic patterns, given the sparsity of the data. A user day normally has 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.

The same spatial and time discretization will be used to define the attributes for this task. Although, before the clustering can be performed, some decisions have to be taken concerning to what values will be used as information to represent the data and what clustering algorithms to use for obtaining the user profiles. All this problems will be addressed in the following subsections.

6.1. *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. In our case we can make the analogy of user events with words and the collected behavior for each user as documents.

To obtain the summary for each user behavior, a feature vector is generated following the vector space model/bag of words (BoW) using the geographical and time discretizations. For the attribute values, we have to compute a term frequency (TF) and inverse document frequency (IDF) (Weiss, Indurkhya, & Zhang, 2010). 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: *Absolute term frequency*, computed as the times the user has been in an area during an specific time interval; *Normalized term frequency*, computed as the times the user has been in an area during an specific time interval, normalized by the total number of areas the user has been; and *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 a total of six different representation of the users data.

6.2. *Clustering algorithms*

Different cluster algorithms can be applied to extract group profiles. Due to the representation of the data (BoW), clustering algorithms usually applied for this representation would be successful in finding meaningful clusters. We experimented with three different clustering algorithms: K-means (Dubes & Jain, 1988), spectral clustering (Ng, Jordan, Weiss, et al., 2002) and affinity propagation clustering (Frey & Dueck, 2007).

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 to decide the correct number of clusters. This task is harder because the assumption of spherical clusters is probably incorrect for the clusters in the data, so the common quality indices used to find the number of clusters will not be very

informative. This arises the need for experimenting with different numbers of clusters and to evaluate other subjective characteristics of the clusters.

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

Spectral clustering is a graph based clustering algorithm that uses the graph Laplacian matrix. The eigenvectors of this matrix are obtained and used as a transformation for the original data that maintains their local structure. This allows to discover non spherical clusters. After the transformation, different algorithms can be used to obtain the clusters, for instance K-means. The computational cost is the same as affinity propagation. Also this algorithm has been applied successfully for datasets with a bag of words representation.

7. Extracting user behavior profiles

In order to enhance the quality of the data, we filtered users without a minimum number of distinct events. This allows to extract more meaningful profiles with the cost of reducing the number of users. Given that the dataset is sparse, a large portion of users has not been captured a significant number of times, so makes sense to discard this information for our purposes. Also, given that the collection period 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. This reduces the number of users to around ten thousand, depending on discretization granularity. We consider this number of users significant enough to show very different profiles.

In the clustering results, certain number of small sized clusters is bound to appear for profiles not very represented in the data. As a quality criteria we have considered that a cluster is significant if it 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.

7.1. *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 obtained with a size larger than the support.

For instance, for Twitter data, given a discretization and using K-means looking for a

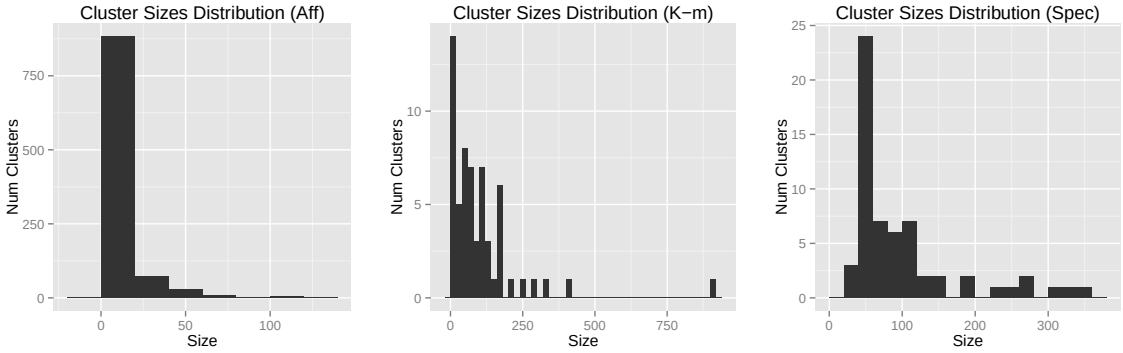

Figure 4. Cluster sizes distribution for Twitter data (two time intervals, 250m diameter) for k-means and spectral clustering with 60 clusters and affinity propagation with damping factor 0.5 using binary frequency attributes.

large number of clusters (60), the absolute term frequency obtains only a cluster with most of the examples and the rest correspond to clusters with less examples than the selected support. For the normalized term frequency, only around four clusters appear above the support, with a cluster having more than two thirds of the examples. Using the binary term frequency results in more than forty clusters with a wide range of cluster sizes. Similar results are obtained with different data discretizations and clustering algorithms with this dataset and also with the Instagram dataset for both cities.

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

7.2. *The clustering algorithms*

Given the different assumptions and bias of clustering algorithms, we need to analyze their results respect to the kind of clusters obtained. In this section, the distribution of the sizes of the clusters and the similarity of the clusterings will be analyzed.

Affinity propagation clustering automatically determines the adequate number of clusters for the data. This only depends on one parameter of the algorithm, the damping factor, that has values in the range 0.5 to 1. With our datasets, using the extreme values for the damping factor and depending also on the discretization of the data, a very large number of clusters is returned, ranging from around 600 to 1100 clusters. A large proportion of this clusters are below the support threshold, leaving with around 75 to 100 not very large clusters. 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 obtained by affinity clustering a range between 60 and 100 clusters was considered. The results show that a large portion of the clusters are under the support as it happens for affinity propagation 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 between 30 and 40 clusters are obtained, for 250m between 45 and 55 clusters and for 500m between 50 and 75 clusters. The distribution of sizes is more reasonable, having a very large cluster including 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 split, remaining the larger clusters intact.

For spectral clustering, we have also used K-means as the clustering algorithm for the post process after the transformation using the Laplacian matrix. The same range of clusters than for K-means was used. The results show that there are almost no clusters

(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								-

Table 2. Cluster similarity among different clustering algorithms using AMI index for Barcelona Twitter data with two time interval discretization and 250m space discretizations.

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 distribution of the sizes of the clusters is less extreme 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 4 shows the distribution of the sizes of the clusters for Barcelona Twitter data. 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.

To measure how much is shared among the clusters obtained with the different clustering algorithm, external validity measures are a useful tool. To measure this, the Adjusted Mutual Information (AMI) index (Vinh, Epps, & Bailey, 2010) has been used. This measure can be used to compare with a reference partition or as a relative measure among different partitions. Its value is in the range [0,1], being 1 identical partitions.

From the experiments, time discretization does not seems to affect much to the similarity among the clusterings. The space discretization increases the similarity among the clusters when is coarser. Table 2 shows the values for AMI measures using two time interval discretization and space discretizations of 250m. From the value of the measure, it looks that the similarity among the clusterings is not high, due to the large number of clusters, specially for affinity propagation. This measure indicates that K-means and spectral clustering results are more similar to each other and equidistant to affinity clustering. 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 further validation.

7.3. *Clusters interpretation*

In order to facilitate the interpretation of the clusters by the expert, a prototype, represented over a map, is computed as the normalized number of visits of the users to the different areas 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 normalized number of visits broken 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 indicated by the cluster validity index, 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

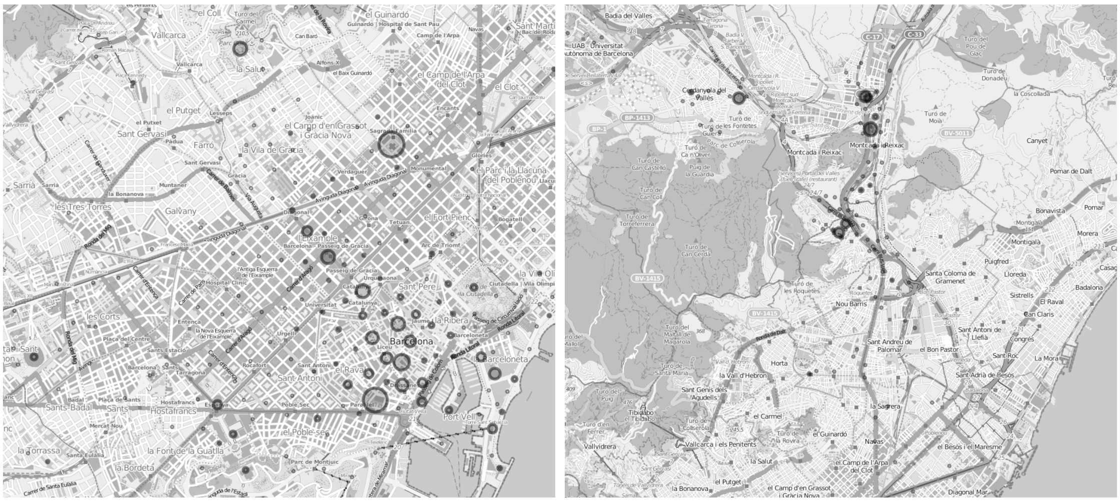

Figure 5. Clusters obtained by K-means for Barcelona Twitter data, a cluster (left) described mainly by touristic points of interest, and a cluster (right) that shows a rush hour pattern where the more frequent places are concentrated along two highways that enter Barcelona from the north-west.

experts that appear only for a particular clustering algorithm.

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.

A more profound knowledge of the domain is necessary to interpret the specific clustering results, 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, 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 nearby 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. Fourth, 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 patterns. Figure 5 shows some examples of these kinds of clusters.

8. 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 from 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 two methodologies that are 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 points of interest or citizens profiles. 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 Twitter events with Instagram events for the same user, reducing this way the sparsity of the data. In the analysis only the geographical position of the events and their timestamp have been used, but other useful characteristics can be included like, for instance, the language or the social connections of the user. 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.

9. Acknowledgments

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

References

Andrienko, N., & Andrienko, G. (2012). A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 1-29.

Dubes, R., & Jain, A. (1988). *Algorithms for Clustering Data.* Prentice Hall.

Eagle, N., & Pentland, A. (2006, March). Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Comput.*, *10*(4), 255–268.

Farrahi, K., & Gatica, D. (2011). Discovering Routines from Large-scale Human Locations Using Probabilistic Topic Models. *ACM T. Intell. Syst. Technol.*, *2*(1), 3:1–3:27.

Frey, B., & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, *315*.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *DM & KD*, *8*(1), 53–87.

Joseph, K., Tan, C. H., & Carley, K. M. (2012). Beyond Local, Categories and Friends: Clustering Foursquare Users with Latent Topics. In *UbiComp 2012* (pp. 919–926).

Lee, R., Wakamiya, S., & Sumiya, K. (2013). Urban area characterization based on crowd behavioral lifelogs over twitter. *Pers. & Ubiquitous Computing*, *17*(4), 605–620.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM.*

Ng, A., Jordan, M., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, *2*, 849–856.

Pianese, F., An, X., Kawsar, F., & Ishizuka, H. (2013). Discovering and predicting user routines by differential analysis of social network traces. In *IEEE 14th Int. Sym. on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (p. 1-9).

Vinh, N., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR*, 2837–2854.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining* (Vol. 41). Springer.

Zheng, Y. (2011). Location-Based Social Networks: Users. In *Computing with Spatial Trajectories* (pp. 243–276). Springer.