Crowdsourcing Land Use Maps via Twitter

Vanessa Frias-Martinez
College of Information Studies
University of Maryland
vfrias@umd.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
efm@tid.es

ABSTRACT

Individuals generate vast amounts of geolocated content through the use of mobile social media applications. In this context, Twitter has become an important sensor of the interactions between individuals and their environment. Building on this idea, this paper proposes the use of geolocated tweets as a complementary source of information for urban planning applications, focusing on the characterization of land use. The proposed technique uses unsupervised learning and automatically determines land uses in urban areas by clustering geographical regions with similar tweeting activity patterns. Three case studies are presented and validated for London (UK) and Madrid (Spain) using Twitter activity and land use information provided by the city planning departments. Results indicate that geolocated tweets can be used as a powerful data source for urban planning applications.

Keywords

Urban Computing;Land Use Modeling;Geolocated Tweets

1. INTRODUCTION

Urban planning focuses on the design of urban environments so as to increase the well being of citizens. In this context, urban planners are interested in understanding how different parts of the urban landscape are being used by citizens. For example, analyzing whether an area is residential or industrial. Urban planners often attempt to gather land use information through questionnaires or in-person interviews. This traditional approach has some limitations such as the cost, which highly limits the frequency with which the information is captured. Alternative approaches such as GIS (Geographic Information Systems) provide satellite imagery that might reveal some types of land use information through image processing techniques [17]. However, such techniques fail to provide real time information as images are not captured frequently and the land uses that can be identified do not cover the variety of land uses present in a city.

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.

The increasing use of ubiquitous and mobile technologies is generating large-scale datasets containing information about how citizens interact with their environments. These data sources are becoming relevant for urban planning applications such as transport planning [6] or traffic estimation [2]. In the area of urban land use, several pervasive technologies have been used to characterize urban behaviors including GPS [18], cell phone traces [14] or social media applications such as Foursquare [10]. In general these approaches tend to focus on a specific location, on specific interactions (*e.g.,* visited places or mobility patterns) and most importantly, they lack a quantitative validation of the results.

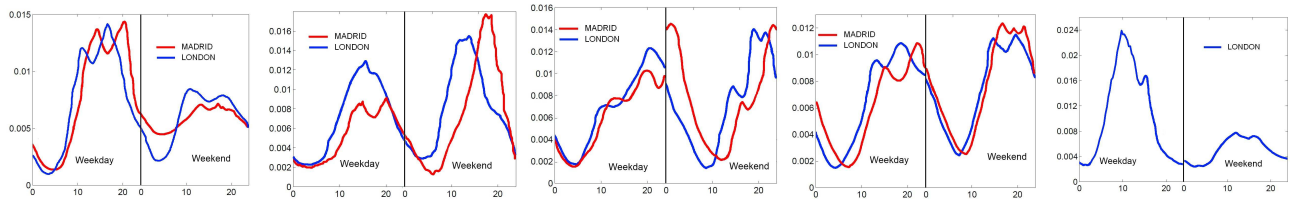
In this paper we propose to use Twitter geolocated data for the automatic identification of land uses. The proposed approach exclusively makes use of spatial (geo-tagged) and temporal (time-stamped) information of tweets, without accessing personal details or the content of the user-generated information. By doing so, it preserves privacy and can potentially be applied and/or complemented with any other mobile social media dataset with geolocation information. Our novel approach is designed to identify all possible land uses using spectral clustering, it is validated using real land use data provided by city planning departments and is tested in two urban environments: London (UK) and Madrid (Spain).

2. SENSING URBAN LAND USES

We propose an unsupervised approach for the automatic identification of urban land uses from geolocated tweets. It consists of two steps: (1) land segmentation, to divide the urban area under study into smaller geographic regions and (2) land use detection, to determine the type of land use for each geographic region.

2.1 Land Segmentation with Geolocated Data

Given that we want to sense land uses in different urban regions, the first step consists on partitioning the land into different segments, which can then be characterized by its usage pattern. The partitioning of the area considered has to preserve the topological properties of the geolocated tweets, while respecting the actual shape of the geographical area under study. We approached this problem using Self-Organizing Maps (SOM) (Kohonen, 1990). We define a SOM consisting of a collection of N neurons organized in a rectangular grid $[p, q]$, with $N = pq$. Since we can choose any initial size $[p, q]$ for the map, our method explores different sizes and selects as the best land segmentation map the topology that minimizes the Davies-Bouldin clustering



(a) Business (b) Leisure (c) Nightlife (d) Residential (e) Industrial

Figure 1: Tweeting activity signatures per cluster for London and Madrid. The Y axis represents the normalized tweeting activity and the X axis two 24-hour periods: weekdays and weekends.

index[5]. Smaller values for the DB index guarantee that the neurons are well separated and that each neuron represents a compact cluster of geolocated tweets. After applying this process, we obtain a map where each neuron represents a pointer to a region with a high density of tweets. We finalize by computing Voronoi tessellation [1] over the set of the neurons (geolocated points) in order to compute the land segments that each neuron represents. The next step will determine the type of land use for each of the land segments (Voronoi polygons).

2.2 Unsupervised Detection of Urban Land Uses

We characterize each land segment by its average tweeting activity which will then be used to identify common land uses. For each land segment s, a tweet-activity vector X_s representing the average tweeting behavior is computed as:

Step 1. An activity vector $x_{s,n}$ for land segment s is built for each day $n = 1, ..., d$ in the dataset.

Step 2. Each day n in the activity vector contains 72 components $x_{s,n}(t), t = 1, ..., 72$ where each one represents the number of tweets generated in land segment s during a 20-minute interval t in day n.

Step 3. An average activity vector for each land segment s is computed for both weekdays $X_{s,wkd}$ and weekends $X_{s,wkn}$, each one representing the average number of tweets in land segment s at each time period t considering only weekdays (Monday through Friday) in the first case and weekends (Saturday and Sunday) in the second: $X_{s,wkd}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}$ and $X_{s,wkn}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}$ with $n = 1, ..., d$ and $t = 1, ..., 72$.

Step 4 The final activity vector is represented as the concatenation of weekday and weekend average activity vectors $X_s = \{X_{s,wkd}, X_{s,wkn}\}$ and is normalized as $\hat{X}_s(t) = \frac{X_s(t)}{\sum_{t=1}^{72} X_{s,wkd}(t) + \sum_{t=1}^{72} X_{s,wkn}(t)}$.

This four-step process allows us to represent each land segment with a unique activity vector X_s that contains 144 elements representing the average weekday and weekend tweeting activity computed in 20-minute timeslots. Next, we apply clustering over these activity vectors to automatically identify and characterize urban land uses. We posit that land use can be derived from a careful analysis of the tweeting behaviors in each cluster, based on its activity vector as well as on its physical layout in the city.

We have selected spectral clustering [12] due to its advantages: no assumptions about the shape of the clusters; ability to manage large dimensional datasets by using dimensionality reduction; easy to implement using standard linear algebra; and generally, good clustering results with a low computational cost. Spectral clustering requires two input parameters: a similarity matrix S that represents the pairwise similarities $s_{i,j} = s(X_i, X_j)$ between all vectors X_k to be clustered as well as the number of clusters k to compute. In our context X_i and X_j represent the tweeting activity vector of each one of the land segments previously obtained. We compute the similarity $s_{i,j}$ as the cosine similarity which assigns values in the range $[-1, 1]$ being one equal vectors and minus one representing exactly the opposite. Regarding the number of clusters k, we use the eigengap detection technique [16] which determines the value of k by the rank of the eigenvalues where there is the largest difference in the value of the eigenvalues of the Laplacian matrix arranged in increasing order. The final output of the spectral clustering are k clusters, each one containing a set of activity vectors. To analyze the type of land use associated to each cluster, we compute an average activity vector that represents the tweeting activity for each cluster. Finally, we hypothesize about the land use for each cluster based on its tweeting activity and its geographical location in the urban environment under study. To validate our results, we contrast our clusters and land uses hypotheses against real land use information collected by the corresponding city planning department.

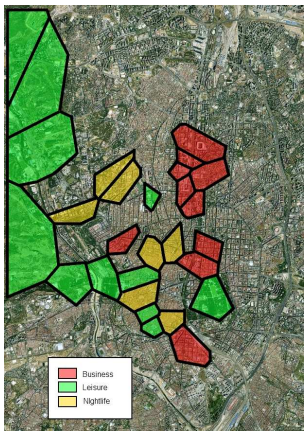
3. EVALUATION OF LAND USES

We present an evaluation of our land use detection method for two metropolitan areas: London (UK) and Madrid (Spain). We have selected these cities because they show different densities of Twitter activity computed as the number of daily tweets per square kilometer in the urban perimeter considered: London has 42.51 tweets/km^2 and Madrid around 10.88 tweets/km^2. As a result, these cities represent different cultural and behavioral Twitter attitudes useful to evaluate the limits of our approach. The objective of this evaluation is to analyze to which extent the land use identification algorithm detects different types of land uses.

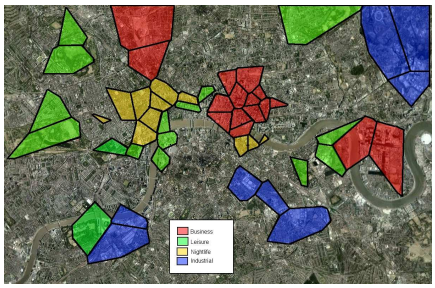
We use the Twitter Streaming API to gather geolocalized tweets in near real-timefor London and Madrid. For London, we collect tweets within the Ringway 1 and for Madrid all tweets comprised within the M30 highway. Approximately about one percent of the full Firehose tweets are geolocalized. Our final Twitter dataset consists of 49 days (seven weeks) of geolocated tweets for London and Madrid.

3.1 Land Segmentation and Land Uses

Our method trains a SOM with the set of geolocated tweets to divide the urban area under study into different land segments s characterized by their tweeting activity vector X_s. Given the different geographies of the cities under study, we evaluated N in the range $N = [10, ..., 300]$. Due



(a) Madrid Land Uses



(b) London Land Uses

Figure 2: Physical layout of business (red), nightlife (yellow), leisure (green) and industrial (blue) clusters. Areas not marked in color are residential.

to the randomized nature of the SOM training stage, 100 SOMs are trained for each city and each pair (p, q) with $N = p * q \in [10, ..., 300]$ and their average DB index is computed. The minimum DB index was N = 168 for London with p = 12 and q = 14 and N = 91 for Madrid with p = 7 and q = 13. Each one of the land segments identified in London and Madrid is characterized by its Twitter activity vector X_s which has 144 components; the first 72 describe the tweeting activity during an average weekday and the last 72 the activity during and average weekend day. Note that the number of X_s vectors for each city is given by the optimal number of SOM neurons identified in each case (168 for London and 91 for Madrid). Our method uses the set of X_s vectors to identify different land uses for each city computing clusters of similar normalized activity using spectral clustering; i.e., the similarity matrix S is computed for each city with the set of corresponding X_s vectors. The best number of land segment clusters is $k = 4$ for Madrid and $k = 5$ for London.

In order to understand the types of land uses identified by these clusters, we analyze the class representatives for each cluster together with its geographical distribution over the city map. A combined analysis can be used to provide a hypothesis about the potential types of land uses. Figure 1 presents the class representatives for each one of the clusters identified across the three cities. Each representative (behavioral signature) is computed as the average number of hourly tweets and is normalized per cluster and per city. For analytical purposes, we group the signatures across cities by Euclidean similarity. We hypothesize that signatures that share similar shapes across cities represent comparable land use types.

We observe that the activity vectors in *Cluster 1* are generally characterized by a larger tweeting activity during weekdays than weekends (see Figure 1(a)). During weekdays the highest tweeting activity is reached at around 10:00AM and 18:30PM for London, which might be associated to the times at which people typically get to work, go for lunch, and leave work. In the case of Madrid, the signature is shifted, suggesting that working hours might happen a little bit later during the day. The peak of the tweeting activity during the weekends is reduced by approximately 40% when compared to weekdays. In terms of geolocation of the clusters, these cover, among others, areas like the City and Canary Warf in London (see Figure 2(b)) and the surroundings of Castellana and the area of AZCA in Madrid (see Figure 2(a)), all areas heavily associated with business/office activities. For these reasons, we hypothesize that the geographical area covered by this cluster represents Business areas in London and Madrid.

Cluster 2 shows a large difference between weekend and weekday activity, in fact, the signature is almost doubled in volume (see Figure 1(b)). During weekends, tweeting activity increases until the afternoon, and constantly decreases after that. Geographically, these clusters cover regions like Hyde Park or Regents Park in London (see Figure 2(b)) and El Retiro Park and Casa de Campo Recreational Park in Madrid (see Figure 2(a)). Also included are heavily touristic areas, like Sol and the Flea Market of El Rastro in Madrid, or the London Eye, Buckingham Palace and Covent Garden in London. Thus, we hypothesize that this cluster can be associated to Leisure or Weekend activities since users are active mostly during the weekends. However, we believe that it does not represent weekend nightlife since the tweeting activity highly decreases after 16:00PM during the weekends.

On the other hand, *Cluster 3* is associated to very large activity peaks at night (see Figure 1(c)). These peaks happen at around 20:00-21:00PM during weekdays and between 00:00-06:00AM during the weekends. We observe that the peaks happen earlier in London while a little bit later in Madrid suggesting that nightlife might continue until late hours in this city. Studying the physical layout of these clusters on the city maps, we observe that they cover areas like the West End in London and Malasana in Madrid (see Figure 2), areas associated with restaurants and clubs. All these elements suggest that this cluster might represent nightlife activities. *Cluster 4* shows a signature evenly divided between weekends and weekdays, where, during weekdays, there is a peak of activity in the afternoon between 6pm and 8pm. Activity during weekends is of the same magnitude as in weekdays (see Figure 1(d)). This is the largest cluster in terms of total area and it covers heavily residential areas in all cities. In Figure 2, the areas include with this cluster are the ones not marked with any color. Our hypothesis for this type of signature is that it represents residential land use with citizens tweeting from home at any time during the weekends and after working hours during the week.

Finally, *Cluster 5* is only identified for London (see Figure 1(e)). Its signature is characterized by a reduced activity during the weekends. The weekdays show a very early peak

Official Land Use	Twitter Land Use				
	Bus.	Resid.	Night.	Leis.	Ind.
London					
Non − domestic buildings	**61%**	9%	3%	2%	25%
Domestic buildings	9%	**56%**	23%	6%	6%
Greenspace&Paths	8%	11%	7%	**72%**	2%
Madrid					
Commercial&Business	**69%**	25%	4%	2%	–
Residential	11%	**61%**	18%	10%	–
Industrial	58%	33%	**3%**	6%	–
Greenspace	7%	16%	6%	**71%**	–

Table 1: Percentage of overlap between official land uses and Twitter land uses for London and Madrid.

in activity (10am), after which decreases for the rest of the day. Looking at the physical layout, these clusters cover areas in the east and south of the city: around Battersea Station and the Olympic Park. Thus, we hypothesize that this cluster represents Industrial land use (see Figure 2(b)). Finally, it is important to clarify that we have only focused on identifying the main land use of each cluster (although there might be other minor ones), since this is the way urban planners typically compute land use maps.

3.2 Land Use Validation

In order to validate our land use hypotheses, we compare the evaluation results against official land use ward profiles released by the London Datastore Open Data Initiative [7] and against the district land use information computed by the Urban Planning Department in Madrid's City Hall [8]. These catalogs are produced by city agencies typically through a combination of on-site inspections, interviews and questionnaires. The information provided by the London Datastore considers three types of wards: (1) domestic buildings, which we associate to residential areas, (2) non-domestic buildings, which we pair up with business and industrial land use wards and (3) green spaces and paths. Finally, the information provided by the City Planning Department in Madrid provides land use information at a district level and considers four types: (1) residential areas with different density levels (which we group), (2) industrial, (3) services(commercial & business) and (4) green spaces.

To understand how well the clusters we have identified using Twitter activity represent the official land use areas, we evaluate the percentage of overlapping that exists between the physical layout of the clusters and the official land use map for each city under study. Such analysis will give us an understanding of the accuracy of our approach to identify land uses as well as of the impact that the Twitter density might have on the quality of the results. It is important to highlight that the percentage of overlapping is an approximate measure to validate land use identification given that both maps have different granularities: our cluster maps represent land segment clusters based on Voronoi and tweet density whereas the official land use maps show data at a ward or district level, depending on the city. Table 1 shows the percentages of overlap between the official land use maps for each city (rows) and our land use hypotheses (columns). Each element (i, j) in the tables represents the percentage of the official land use region that is covered by one of our land use clusters i.e., Business, Residential, Nightlife, Leisure and Industrial.

The official Commercial and Business land uses are identified quite well by our business cluster with area coverage between 61% − 81%. London is a special case in which the official non-residential land use is partially identified by our business cluster (61%) but also by our industrial cluster (25%). Similarly, the official Residential/Domestic buildings land use has a high overlap with our residential cluster with coverage between 56% and 68% of the official areas. However, we observe a generalized trend across the two cities whereby around a 20% of the official residential area is also covered by our nightlife clusters, probably highlighting residential areas with high densities of bars and restaurants. This is in fact common in areas like Chelsea in London or Chueca in Madrid. While in London we are able to detect Industrial land use, and compare it to the official non-residential land use, the official Industrial land use, present in Madrid, goes undetected. We consider that the main reason for that is that, within the area of the city considered, industrial land is minimum (less than 3% of the total area in Madrid), and as a result they are included in larger Voronoi elements that has a different stronger land use. In fact, most of the official industrial land use is subsumed by our business cluster. This might indicate that workers in the industrial areas are not using Twitter as much as people that live and/or work in that area, and as a result our technique captures the main land use, i.e. the official land use goes undetected due to lack of activity. Finally, the official Parks & Recreation and Greenspace & Paths land use is identified by our leisure cluster with overlaps between 71% and 81% of the official land use maps.

Our evaluation and validation for two different cities with varied physical layouts shows two important results. First, our methodology constitutes a good complement to model and understand in an affordable and near real-time manner land uses in urban environments. In fact, we have shown that residential, commercial and parks & recreation areas are well identified with coverage above 70%. Also, our approach is able to identify a land use, nightlife activity, which is not currently modeled by city halls. This has implications from a planning perspective as these areas usually cause noise and security problems and can move over time.

4. RELATED WORK

LBS and social media has been used in the field of urban computing: Noulas et al. [10] and Cranshaw et al. [4] have used the geolocated information provided by Foursquare to model crowd activity patterns and social dynamics; and Neuhaus [9] presented preliminary results on using Twitter for characterizing urban landscapes. As for CDRs, Soto et al. [13], Calabrese et al. [3], Ratti et al. [11] and Toole et al. [15] have used cell-phone records to characterize individual and crowd patterns in urban environments.

5. CONCLUSIONS

In this paper we have presented and validated an unsupervised approach for identifying land uses using location-based social media in London and Madrid. The results have shown that geolocated tweets can constitute a good complement for urban planners to model and understand traditional land uses (like industrial or residential) and identify new ones (like night activities) in an affordable and near real-time manner.

6. REFERENCES

[1] F. Aurenhammer. Voronoi diagrams a survey of a fundamental geometric data structure. 23(3):345–405, 1991.

[2] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. Del Castillo. Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1430–1441, 2012.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[4] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *International Conference on Weblogs and Social Media, ICWSM*, 2012.

[5] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[6] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez. Estimation of urban commuting patterns using cellphone network data. In *In Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 9– 16, 2012.

[7] London. London open data. `http://data.london.gov.uk/visualisations/atlas/ward-profiles-summary/atlas.htm`.

[8] Madrid. Madrid open data. `http://www.madrid.org/cartografia/idem/html/web/index.htm`.

[9] F. Neuhaus. New city landscape: Mapping urban twitter usage. *Technoetic*, 9(1):31–48, 2011.

[10] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *3rd Workshop Social Mobile Web (SMW), International Conference of Weblogs and Social Media*, 2011.

[11] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[12] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[13] V. Soto and E. Frias-Martinez. Automated land use identification using cell-phone records. In *The 3rd ACM International Workshop on Hot Topics in Planet-Scale Measurement (HotPlanet)*, 2011.

[14] V. Soto and E. Frias-Martinez. Robust land use characterization of urban landscapes using cell phone data. In *The First Workshop on Pervasive Urban Applications (PURBA)*, 2011.

[15] J. L. Toole, M. Ulm, M. C. Gonzalez, and D. Bauer. Inferring land use from mobile phone activity. In *In Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 1–8, 2012.

[16] U. Von Luxburg. A tutorial on spectral clustering. statistics and computing. 17(4):395–416, 2007.

[17] J. Yin, Z. Yin, H. Zhong, S. Xu, X. Hu, and J. Wu. Monitoring urban expansion and land use/cover changes of shangai metropolitan area during the transisional economy (1979-2009) in china. In *Environmental Monitoring and Assessment*, volume 177(1-4), pages 609–621. 2011.

[18] Y. Yuan, J.and Zheng and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194, 2012.