Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks

Anastasios Noulas
Computer Laboratory
University of Cambridge
an346@cl.cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge
ss824@cl.cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cl.cam.ac.uk

Massimiliano Pontil
Computer Science Department
University College London
m.pontil@cs.ucl.ac.uk

Abstract

Location-Based Social Networks (LBSN) present so far the most vivid realization of the convergence of the physical and virtual social planes. In this work we propose a novel approach on modeling human activity and geographical areas by means of place categories. We apply a spectral clustering algorithm on areas and users of two metropolitan cities on a dataset sourced from the most vibrant LBSN, Foursquare. Our methodology allows the identification of user communities that visit similar categories of places and the comparison of urban neighborhoods within and across cities. We demonstrate how semantic information attached to places could be plausibly used as a modeling interface for applications such as recommender systems and digital tourist guides.

Introduction

The thriving rise of Location-based Social Networks (LBSN) driven by the increasing adoption of smartphone devices is brining a new set of opportunities for research scientists and application developers. A novel characteristic of those emerging data sets is the fact that they are not generated in a controlled experimental context, but instead, *anyone* can participate and share data at will. While previous mobility data featured GPS geographic coordinates of users, or their estimated position via cellular data, LBSNs offer the opportunity to source mobility data with fundamentally different attributes. First, location broadcasts are *focused* at places: when users report their geo-coordinates a specific venue entity, such as a restaurant or a football stadium, is also identified. Moreover, LBSN places are *semantically enriched* with annotations such as place categories, tags, tips or user comments. Finally the *scale* of LBSN data at all dimensions is dependent on user participation.

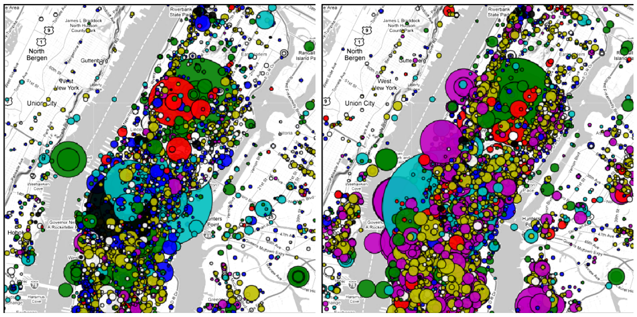
In this position paper, we show how semantic information about places and social activity observed at those could be exploited in the context of future mobile applications and scientific research. In particular, we propose *the use of place categories to create fingerprints of users and areas*: people can be profiled according to the types of places they visit, whereas geographic areas can be modelled according to their constituent venues. Using unsupervised learning techniques (Shi and Malik 1997), we demonstrate how this representation enables the identification of clusters of geographic areas and users within two metropolitan cities, London and New York, and we discuss similarities and differences of findings between them.

Our work is based on a large data corpus obtained by publicly available Foursquare location broadcasts. We describe the obtained corpus at the next section, followed by a description of our clustering methodology and results. We conclude with a discussion on potential future steps and applications.

Foursquare Dataset

Users can use the Foursquare application on their mobile devices and when at a place they can *checkin*, letting their friends or the *world* (if they edit their privacy settings accordingly) know where they are. These checkins can be further pushed to other social networking platforms such as Twitter and Facebook. Foursquare has a game element integrated, allowing users to become *mayors* of a place, if they have the highest number of checkins in the last sixty days.

Since Foursquare data is not publicly available and the corresponding API provides rate limited access, we have resorted to another channel to collect publicly available Foursquare data: Twitter messages which contain Foursquare checkins. Through the public stream of Twitter messages, or tweets, we have recorded approximately 12 million timestamped location checkins, generated by 679 thousand Foursquare users, between May, 27th 2010 and September, 14th 2010. Each message corresponds to a checkin at one of the 3 million recorded locations on the planet. A spatio-temporal margin of the collected corpus is depicted Figures 1(a) and 1(b), where user activity in New York is depicted for morning and night respectively. A circle represents a venue and its radius the popularity of it in terms of number of checkins. Each color corresponds to one of the 8 general categories introduced by Foursquare (described in caption). The *mosaic* created by user checkin data highlights the diversity of human activity across the spatial plane. Next, we demonstrate how we make use of it to represent and cluster geographic areas.



(a) Morning (b) Night

Figure 1: New York at morning and night. We depict the 8 eight categories of the top hierarchical level: Arts & Entertainment (red), College & Education (black), Shops (white), Food (Yellow), Parks & Outdoors (green), Travel (cyan), Nightlife (magenta), Home/ Work/ Other (blue).

City and User Clustering

In this section we model activity patterns across the geographic areas of two metropolitan cities, London and New York, and Foursquare users active in them. We apply a clustering algorithm and present a series of experimental results aimed at answering questions such as *How similar are two areas of a city?* or *Can we find areas between New York and London that may resemble each other?* Similarly, switching the perspective onto users, we would like to compare different users in Foursquare and decide how much their activities look alike. To do this, we position Foursquare categories at the core of the clustering methodology. This is achieved by their use as *features* to represent geographic areas and users. We represent an area according to activity at nearby places. In a similar manner, a user's activity is represented based on the types of places she visits.

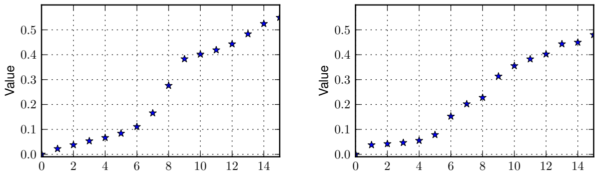
Representation of Geographic Areas Our methodology for the representation of geographic areas is the following: we consider a centre point g within a city and a large square area A. We split A into a number of equally sized squares, each one representing a smaller local area a. Each area a will be a datapoint input for the clustering algorithm. The representation of a is defined according to the categories of *nearby places* and the attached *social* activity modelled through the number of checkins that took place at those. In this way not only we know what types of places are in an area, but we also have a measure of their importance from a social point of view. According to the above notions, we define the signal, $cs_{c,a}$, of a category c to a geographic area a, for all places p that belong to category c within a, as follows:

$$cs_{c,a} = \sum_{p \in c} \#\text{checkins(p)}, \quad \forall p \in a. \tag{1}$$

Hence, each area a can be represented using a vector $\mathbf{cs_a}$, the dimensionality of which is the number of different categories and each feature's value is equal to the signal (normalized over total checkins in the area) received from a particular category. We define the similarity between two areas a and b as the cosine similarity between their corresponding feature vectors:

$$sim_{a,b} = \frac{\mathbf{cs_a} \cdot \mathbf{cs_b}}{||\mathbf{cs_a}||\,||\mathbf{cs_b}||} \tag{2}$$

Having defined the similarity equation between the input data points, we can now create the weight matrix W and the degree matrix D that will be utilized by the Spectral Clustering algorithm (Shi and Malik 1997). The algorithm performs a non-linear dimensionality reduction on the inputs and then a k-means algorithm is applied to the low-dimensional embedding. At the heart of the algorithm lies the *Graph Laplacian L*, where $L = I - D^{-1} - W$. To decide the parameter k that corresponds to the number of output clusters, we have used the *eigengap heuristic* (Luxburg 2006). The latter suggests that one should detect the largest difference between two consecutive eigenvalues of the Laplacian matrix L and set k equal to the rank of the eigenvalues. For instance, in Figure 2, k is 8 and 9 for New York and London respectively.



(a) New York Areas (b) London Areas

Figure 2: Eigenvalue Distribution of Graph Laplacian

Area Clustering Results According to the above formulations we have elected two center points g, encoded as latitude and longitude pairs (51.513,-0.117) for London and (40.764,-73.979) for New York. The large areas A are set to cover a region of $10{\times}10$ km^2. An important parameter to consider is the square size of the smaller areas a, which are represented according to Equation 1. The tradeoff here involves choosing a small square size to favor the characterization of specific areas, but with a large enough number of checkins to obtain a statistically sound representation. Imposing a threshold of at least thirty checkins per area has yielded 228 areas for London and 214 New York respectively, with corresponding sizes $625{\times}625$ m^2.

We now demonstrate the results yielded by the clustering algorithm. Each cluster is represented via its centroid with the top five features ranked according to their popularity amongst the cluster members. A common observation for both cities as seen in Table 1, is the fact that areas have a dominant feature, usually accompanied by a second that is also highly popular. At the case of New York, Cluster 2 may signify residential areas, Cluster 1 outdoor areas and parks (in Figure 3 covers Times Square, Central Park and Rockefeller State Park Preserve), whereas Cluster 5 suggests the coexistence of Art and Entertainment areas nearby parks,

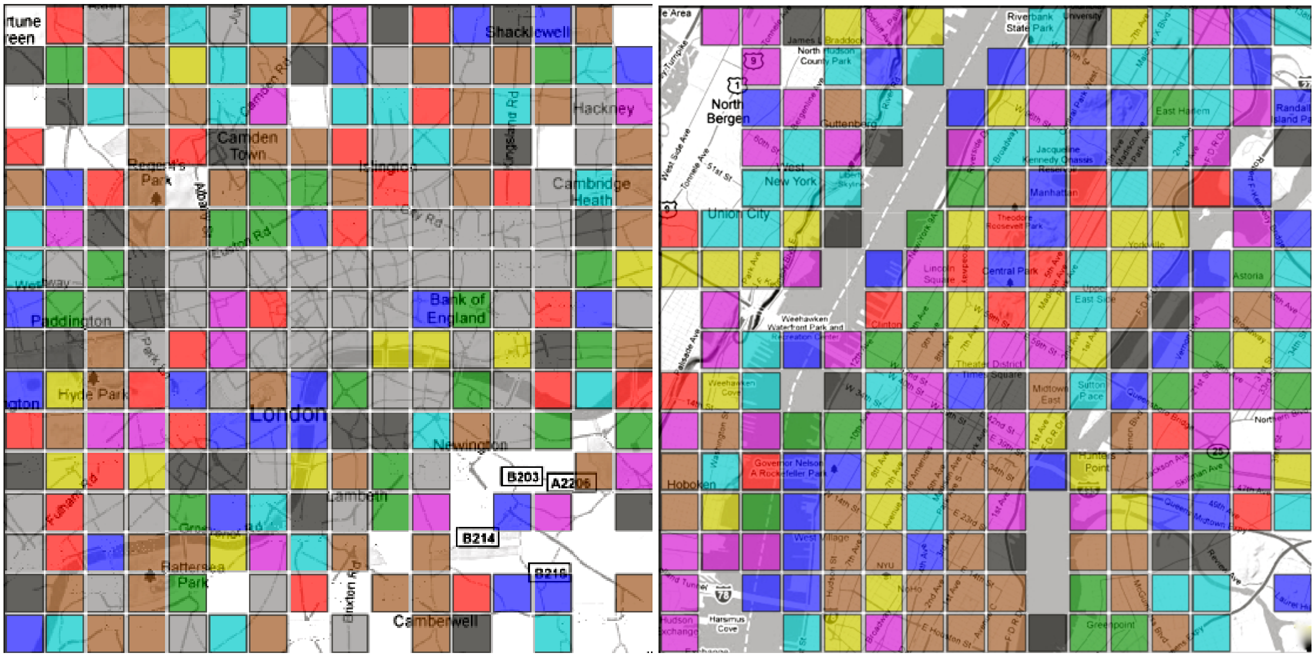


Figure 3: Spectral Clustering visualisation. London (left) and New York (right). Cluster 1: Blue, Cluster 2: Black, Cluster 3: Cyan, Cluster 4: Yellow, Cluster 5: Red, Cluster 6: Green, Cluster 7: Magenta, Cluster 8: Brown, Cluster 9 (London): Gray

Cluster 1(33)	Cluster 2(15)	Cluster 3(35)	Cluster 4(27)	Cluster 5(13)	Cluster 6(16)	Cluster 7(42)	Cluster 8(33)
Parks (0.77)	Home (0.66)	Food(0.57)	Food(0.37)	Arts (0.55)	Nightlife(0.61)	Travel(0.36)	Food(0.42)
Home (0.05)	Parks (0.09)	Shops(0.13)	Shops(0.17)	Parks (0.17)	Food(0.1)	Shops(0.31)	Nightlife(0.31)
Nightlife(0.05)	College (0.06)	Home (0.1)	Home (0.12)	Food(0.07)	Home (0.09)	Food(0.12)	Shops(0.09)
Shops(0.04)	Travel(0.05)	Travel(0.06)	Nightlife(0.1)	Shops(0.07)	Arts (0.05)	Home (0.07)	Home (0.05)
Food(0.03)	Food(0.04)	Parks (0.05)	Travel(0.09)	Home (0.06)	Shops(0.05)	Nightlife(0.05)	Parks (0.05)
Other(0.06)	Other(0.1)	Other(0.09)	Other(0.15)	Other(0.08)	Other(0.1)	Other(0.09)	Other(0.08)

Cluster 1(21)	Cluster 2(23)	Cluster 3(20)	Cluster 4(10)	Cluster 5(24)	Cluster 6(19)	Cluster 7(12)	Cluster 8(45)	Cluster 9(54)
Home (0.54)	Travel(0.37)	Nightlife(0.46)	Arts (0.54)	Food(0.48)	Travel(0.69)	Shops(0.62)	Nightlife(0.44)	Nightlife(0.32)
Food(0.16)	College (0.14)	Travel(0.26)	Travel(0.14)	Nightlife(0.2)	Shops(0.11)	Nightlife(0.11)	Parks (0.29)	Food(0.25)
Travel(0.09)	Nightlife(0.14)	Food(0.11)	Food(0.08)	Travel(0.09)	Nightlife(0.06)	Food(0.1)	Home (0.06)	Home (0.11)
Nightlife(0.08)	Food(0.13)	Home (0.06)	Nightlife(0.07)	Shops(0.06)	Food(0.05)	Travel(0.06)	Food(0.05)	Parks (0.09)
Parks (0.04)	Home (0.09)	Parks(0.03)	Parks (0.06)	Arts(0.04)	Home (0.04)	Home (0.04)	Travel(0.04)	Shops(0.09)
Other(0.09)	Other(0.13)	Other(0.08)	Other(0.11)	Other(0.13)	Other(0.05)	Other(0.07)	Other(0.12)	Other(0.14)

Table 1: Area Clustering. New York (Above), London (Below). Some categories are abbreviated (eg. Home/Work/Other is Home and Parks & Outdoors is Parks).

as also captured by the four red squared areas next to Central Park. Category Food is dominant at Clusters 3, 4 and 8 with the latter case corresponding to areas that also feature Nightlife options. Cluster 7 is a representative of areas with Travel spots, such as metro stations and shops, and has 42 area points which is the highest membership score amongst all clusters. Unlike New York, where Food is the top ranked feature in three clusters, London's principal characteristic in most areas is Nightlife, as highlighted by Clusters 3, 8 and 9. Each of those clusters contains a different secondary feature, namely Travel, Parks (also seen over Regent's Park and Hyde Park areas) and Food. It is notable to observe the latter cluster in Figure 3 (left), covering a wide and geographically cohesive area at the centre and centre east of London. Similar pattern but to smaller spatial extent is observed in New York (see Cluster 8), also with the two top ranked features being Food and Nightlife. Finally, it is interesting to observe that London's Cluster 6 reveals the principal travelling gateaways of the city.

User Clustering Results As with geographic areas, we have represented users according to the types of places they checkin and the checkin frequency in those. Our goal here is to profile Foursquare users and detect groups of individuals with similar activity patterns. We focus on user activity in the two cities by considering *local active* users defined as those whose majority of checkins have been observed within the $10 \times 10\ km^2$ areas (*local*) and the total number of checkins is above 30 (*active*). The filtering has yielded 2433 users for New York and 540 for London.

We now present the clustering results obtained for the Foursquare users of New York and London in Tables 2(a) and 2(b). As with areas, user clustering results are characterised by clusters that have two or three principal features.

Cluster 1(268)	Cluster 2(479)	Cluster 3(202)	Cluster 4(339)	Cluster 5(701)	Cluster 6(223)	Cluster 7(221)
Home (0.51)	Food(0.34)	Nightlife(0.53)	Food(0.41)	Food(0.26)	Shops(0.37)	Food(0.62)
Food(0.15)	Nightlife(0.29)	Food(0.19)	Nightlife(0.13)	Home (0.23)	Food(0.26)	Shops(0.11)
Nightlife(0.1)	Arts (0.08)	Home (0.07)	Shops(0.12)	Nightlife(0.12)	Home (0.08)	Nightlife(0.08)
Shops(0.07)	Shops(0.07)	Shops(0.05)	Home (0.08)	Shops(0.09)	Nightlife(0.08)	Home (0.04)
Travel(0.05)	Home (0.06)	Travel(0.04)	Parks (0.07)	Travel(0.08)	Travel(0.06)	Travel(0.03)
Other(0.12)	Other(0.16)	Other(0.12)	Other(0.19)	Other(0.22)	Other(0.15)	Other(0.12)

Cluster 1(62)	Cluster 2(95)	Cluster 3(72)	Cluster 4(66)	Cluster 5(95)	Cluster 6(119)	Cluster 7(31)
Nightlife(0.61)	Nightlife(0.35)	Home (0.35)	Food(0.58)	Home (0.59)	Food(0.22)	Travel(0.62)
Food(0.18)	Food(0.32)	Nightlife(0.29)	Nightlife(0.13)	Food(0.12)	Travel(0.18)	Food(0.12)
Home (0.06)	Arts (0.09)	Food(0.18)	Shops(0.08)	Travel(0.11)	Shops(0.17)	Home (0.1)
Travel(0.04)	Home (0.07)	Travel(0.07)	Home (0.06)	Nightlife(0.08)	Home (0.14)	Nightlife(0.06)
Parks (0.03)	Travel(0.06)	Shops(0.05)	Travel(0.06)	Parks (0.04)	Nightlife(0.14)	Shops(0.05)
Other(0.08)	Other(0.11)	Other(0.06)	Other(0.09)	Other(0.06)	Other(0.15)	Other(0.05)

Table 2: User Clustering. New York (Above), London (Below). Some categories are abbreviated (eg. Home/Work/Other is Home and Parks & Outdoors is Parks).

With respect to user and area clustering parallels, there is a pattern concerning the appearance of *Food lovers* for New York and *Nightlife people* for London. This is not a surprise, since both user and area representations are sourced from the same *human activity processes*. Interestingly, we also find similar pattern of user behaviours between the two cities. For instance, users who are members of Cluster 3 of New York and Cluster 1 of London, demonstrate identical activity behaviours. Not only the ranking of the cluster features is very similar to each other, but also the respective average values are close to each other (Nightlife, Food, Home). A similar case is observed for Clusters 1 (New York) and 5 (London) of the two cities, with users broadcasting more than 50% of their locations mainly from places that belong to category Home/Work/Other. A user profile however, that is not being observed in both cities, is the existence of a group of 223 users in New York whose principal check-in location is at category Shops. Also in the case of area clustering, the Shops feature is not highly ranked for the London case. That may reflect the existence of a fundamentally different shopping culture between the two cities and we consider the investigation of such cultural variations and similarities across the globe a unique opportunity offered by LBSN data.

Discussion and Future Work

As shown in the previous section, profiling users and dividing them in communities is possible. While prior works in social networks have considered the allocation of users to groups through co-authorship networks, user-indicated friendships or cellular data (Newman 2006; Palla, Barabasi, and Vicsek 2007), we are now able to relate mobile phone users with specific places and their respective categories. The opportunity to offer personalized recommendations when a user is on the move based on her past activity behaviour has become possible. In addition, social scientists could acquire a new grasp on users as their activity and likes are being embedded on the digital space and shared over publicly available channels.

The modelling and clustering of geographic areas according to nearby places and user checkins proposes a new way to view the physical space. Fingerprinting areas under those terms, could help the development of new applications. As examples, we envision digital tourist guides (Hao et al. 2010) that could allow the comparison of areas at distant places in the world. User profiles could also be matched to geographic areas at a global scale. Urban scientists could learn about the different types of activities performed across the neighbourhoods of a city, improving past works utilizing urban networks of streets (Porta, Crucitti, and Latora 2006). All this while the continuous participation of users in location-based systems presents a natural way for information to be updated in real time.

In terms of future work we intend to extend the clustering techniques presented here to a number of dimensions. Temporal variations of geographic user activity can be taken into account in order to characterise area and users at certain periods of a day (i.e., morning, night etc.). Moreover, additional semantic information such as topics discussed at areas could be mined by data sourced from user tips, tags and comments. Hence, while a category of a place presents a general characterisation of it, natural language can allow the division of more meaningful representations. Finally, we would like to scale up our work and provide global coverage, including a large number of Foursquare popular metropolitan areas across the planet.

References

Hao, Q.; Cai, R.; Wang, C.; and Zhang, L. 2010. Equip tourists with knowledge mined from travelogues. In *Proceedings of WWW'10.*

Luxburg, U. v. 2006. A tutorial on spectral clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics.

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23):8577–8582.

Palla, G.; Barabasi, A.-L.; and Vicsek, T. 2007. Quantifying social group evolution. *Nature* 446(7136):664–667.

Porta, S.; Crucitti, P.; and Latora, V. 2006. The network analysis of urban streets: A dual approach. *Physica A* 369:853–866.

Shi, J., and Malik, J. 1997. Normalized cuts and image segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition.*