

You Shall Know a Place by the Conversations it Seeds

Syed Fahad Sultan
Technology Innovation Center
Wadi Makkah
Makkah 21955, Saudi Arabia
sfsultan@gistic.org

Hicham G. Elmongui
Comp. & Sys. Engineering
Alexandria University
Alexandria 21554, Egypt
elmongui@alexu.edu.eg

Sohaib Ahmad Khan
Science and Technology Unit
Umm Al-Qura University
Makkah 21955, Saudi Arabia
saashfaq@uqu.edu.sa

Abstract—In this work, we look at problems of urban sensing from the lens of conversations (tweets) on Twitter. Using techniques from statistical natural language processing on geo-tagged tweets, we identify areas which exhibit similar aggregate behavior, infer the land-use of areas and predict types of individual establishments. We demonstrate our inferences using over two years of Twitter data, for a wide variety of spatial contexts and evaluate our results against existing open data sets. Our results are novel in extremely detailed resolution of their mapping, and demonstrate that tweets can be a very effective urban sensor and in many regards are superior to other data sources for studying urban spaces. Our techniques are language agnostic, and can be applied to any city where enough similar data is available.

I. INTRODUCTION

With 54% of the world’s population now living in urban areas [1], a proportion that is expected to increase to 66 per cent by 2050, there is a dire need to better understand the ever increasing complexities of modern cities. Traditional approaches to understanding the city encompass cumbersome survey activities.

In this paper, we tackle problems of urban sensing from the similarities and differences in conversations (tweets) on Twitter, in relation to space. We accomplish this by using a simple yet effective model that we call Space-as-Documents. The Space-as-Documents applies the maxim ‘You shall know a word by the company it keeps’ [2] from statistical language processing to problems of studying cities using text features. We argue and demonstrate that when space is modeled as text documents, techniques from document processing can be applied effectively to many different problems of urban sensing.

Using text of over two years of geo-tagged tweets from New York City, we first identify similar areas in the city (Figure 1), modeling it as a documents clustering problem. Next, we infer land use of different areas in the city, effectively modeling

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

© 2017 Association for Computing Machinery.

ACM ISBN 978-1-4503-4993-2/17/07/\$15.00

<http://dx.doi.org/10.1145/3110025.3110123>

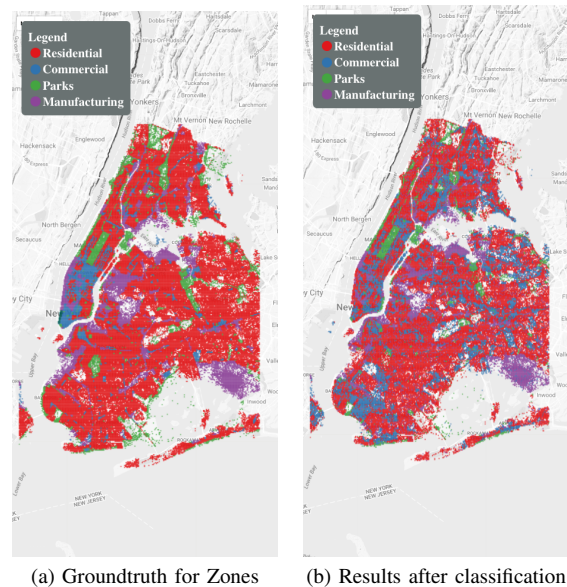


Fig. 1: Land-use categories of New York City: (a) Ground truth for residential, commercial, manufacturing and parks categories. (b) The same categories inferred from Twitter conversations through supervised learning.

it as a multinomial document classification problem, and evaluate against data made public by the local city government. Finally, we use the same techniques to infer the types of points of interest in the city. Here we use points of interest data from Open Street Maps [3] as labels for classification. To demonstrate the effectiveness of our techniques for other highly divergent languages and contexts, we also report results of points-of-interest sensing for Mecca, Saudi Arabia. For each experiment, we present a quantitative evaluation of our results against existing open data sets.

II. SPACE AS DOCUMENTS

Much of modern advances in natural language processing and information retrieval are statistical techniques based on the old adage: ‘You shall know a word by the company it keeps’, decreed by the famous linguist J.R.Firth in the 1950s. Simply put, it states that similar words appear in similar context and vice versa.

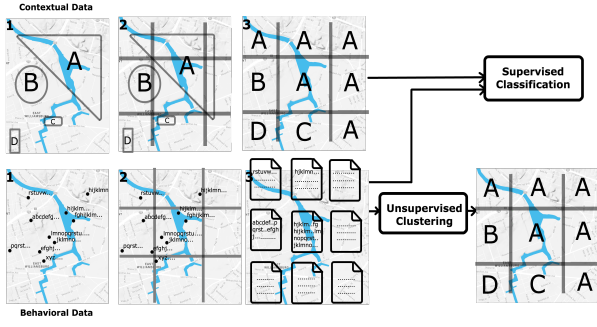


Fig. 2: Approach overview: Behavioral and Contextual data is mapped on to the same spatial structure. Text documents alone, rather than any spatial parameter, act as features for supervised and unsupervised learning.

In addition, from cognitive psychology, we know that context influences perception, and resultantly, behavior. This phenomenon is generally referred to as the Context Effect. It suggests that observed behavior contains key insights into the context it takes place in, or that similar behaviors indicate similar contexts.

The Space-as-Documents model we use here connects these two ideas by modeling space as a set of text documents. Doing so enables us to leverage the amazing progress in recent years in natural language processing and information retrieval for spatial analysis. The Space-as-Documents model is particularly well-suited for problems of urban sensing where it is required to hone in on the relative similarities and differences between different areas in the city.

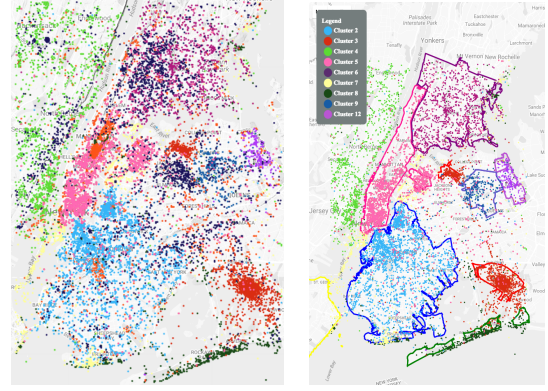
Given geo-located text data D available for N_D spatial units $\{d_1, d_2 \dots d_{N_D}\}$ where $d_i = (location_i, text_i)$ and $location_i, 0 < i < N_D$ is either a polygon or a point, the Space-as-Documents model involves the following three steps:

- 1) **Segmentation** of space using a spatial structure S constituent of N_s spatial units $\{s_1, s_2 \dots s_{N_s}\}$ such that $\forall d_i \in D, \exists s \in S$ such that s spatially contains d_i .
- 2) **Aggregation**: $\forall s \in S$, creation of a representative text document T_s , such that $T_s = aggregate_function(D_s)$ where $D_s \subset D$ and $\forall d_s \in D_s, s$ spatially contains d_s .
- 3) **Representation**: $\forall s \in S$ representation of T_s in feature space.

III. URBAN SENSING USING SPATIAL DOCUMENTS

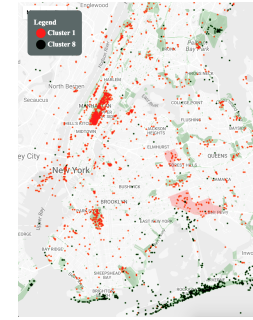
In this paper, we present three experiments on urban sensing: one using unsupervised learning and two using supervised learning. The general approach taken for these is outlined in Figure 2. Behavioral and contextual data sets are both mapped onto the same spatial structure.

Spatial textual data points of behavioral data are aggregated over time and space as detailed in Section II. For all our experiments, we use simple concatenation as the aggregation function to create representative documents. Contextual data is mapped onto the spatial structure such that each unit in the spatial structure is assigned context value of the spatial unit

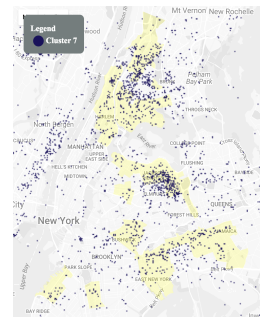


(a) Top twelve clusters from Spectral Clustering on text documents

(b) Nine of twelve clusters with boundaries of corresponding boroughs and neighborhoods



(c) Cluster No. 2 and Cluster No. 8 and parks in the city.



(d) Cluster 10 and neighborhoods with large Hispanic populations.

Fig. 3: Results of clustering for New York City are shown

that makes up the greatest percentage of its area. It is pertinent to point out here that only text based features were used in our experiments and space was not used as a feature.

Document Representations – For our experiments, we use a distributed representation of documents based on Word2Vec [4] called Paragraph Vector [5].













Twitter Data – For our experiments, we used text from geo-tagged tweets for the cities of New York and Mecca, Saudi Arabia collected from October 2013 to March 2016.

A. Unsupervised Urban Sensing

We start off by using unsupervised clustering to discover similar areas in the city. We segment New York City into a grid of equal sized contiguous cells, each cell with dimensions roughly of 110 by 85 meters. Each cell was represented by the text of all tweets from its area, over time, coalesced into a single representative document. Next, Doc2Vec representations of these documents were created. For clustering, we used Spectral Clustering [6].

Figure 3a shows the result. Table I gives cluster numbers and the most important features (words), taken as the vectors closest to the centroid, for each cluster. Each dot in Figure 3a is a cell in our grid. The sparse areas in the city either did not have enough tweets from within them or were associated

TABLE I: Most important features, derived from spectral clustering of text documents

No.	Color	Most important features
1		parks central prospect beautiful walk
2		brooklyn williamsburg bridge bar thank
3		jfk flight airport plane land terminal
4		nj jersey hoboken secaucus hackensack
5		job hire career arc retail veteran apply sale
6		bronx zoo pelham south harlem que life
7		island staten coney ferry long governor
8		beach rockaway orchard brighton far
9		flush meadow queen corona park scan
10		que la el es en como pero esta cuando
11		museum art metropolitan american nature
12		bay terrace side boulevard bell target

with a cluster that had a nearly uniform distribution over the city and was thus removed.

Most clusters in Figure 3a correspond to a unique borough or neighborhood of the city. Figure 3b shows these clusters with polygon boundaries of their corresponding borough or neighborhood. Clusters 2, 5, 7 and 6, correspond to boroughs Brooklyn, Manhattan, Staten Island and Bronx respectively. Most other clusters in Figure 3b correspond to neighborhoods within the fourth borough: Queens. Clusters 3, 8, 9, 12 and 4 correspond to the two airports in the city, Far Rockaway Beach, Flushing, Bay Terrace/Bayside and New Jersey across the Hudson River respectively.

It is important to keep in mind here that space was not used as a feature. Yet still, these clusters exhibit strong spatial patterns. A part of the reason for that lies in Table I. It is evident that in almost all cases, the most discriminable features for a cluster are references to names of places. This is however not true for Cluster 5, the most important features for which are activity-related words, and Cluster 10, which is composed of tweets in Spanish.

Figure 3c focuses on cells of Cluster 1. In the figure, the green polygons show parks in the city. It can be seen that while the cluster includes most major parks in the center of the city, it does not include cells for some parks, cells for which are associated with Cluster 8. Furthermore, there is a high concentration of Cluster 1 cells in some areas, where there are in fact no parks; these are highlighted by red polygons.

It turns out that red polygons are neighborhoods of New York City that have the word 'Park' in their names: Rego Park, Ozone Park and South Ozone Park. On the other hand, cluster 8 corresponds mainly to areas along the coast and cells for some parks are included in Cluster 8 because of their close proximity to water bodies and the resultingly high concentration of words like 'beach' in tweets from them.

These two examples point to the some of the shortcomings in using text for urban sensing.

Figure 3d focuses on cluster 10. As also evident from Table I, this cluster is composed of cells where most tweets are in Spanish. In Figure 3d, the yellow polygons are neighborhoods of New York City with a population of 15000 or more people,

of age 5 and over, who speak Spanish as per the American Community Survey (2009-2013). As it can be seen from Figure 3d, areas with a higher concentration of Cluster 7 cells, correspond quite strongly to these neighborhoods. This is good example of the power of using text for urban sensing. Most other data features conventionally used for urban sensing, such as call detail records or temporal features from social networks, can not provide insights into the distribution of ethnic populations in the city.

B. Land-Use Sensing

Next, we infer land-use of areas for the New York City, modeling it as a multinomial document classification problem. We use Space-as-Documents model to create aggregated documents, representative of different areas in the city and then classifying them against land-use labels.

Land-use data – We obtained ground-truth land-use data from New York City’s Department of City Planning [7]. This data contains 5469 units each belonging to one of the 157 unique land-use categories. Each type further belongs to one of the five broader categories: Residential, Commercial, Park, Manufacturing and Battery Park City. In this study, we only use four land-use types as our labels. We did not include the Battery Park City because areas of the category were too few in number.

For land-use sensing, we segment space using a grid with cells of equal size, each roughly 110x85 meters in their dimension. On mapping land-use categories to cells of our grid, we ended up with 39,012 cells for residential, 7536 for manufacturing, 4970 for Parks, 3506 cells for Commercial and 42 cells for Battery Park City category.

Before creating document representations, all tweets were made to undergo the standard pre-processing steps of tokenization, case-normalization, stop-words removal and stemming. Furthermore, URLs and twitter-handles were excluded from all tweets, since these are not influenced by the spatial context. After pre-processing, we create Doc2Vec representations of documents. For classification, we use a classifier based on Support Vector Machines (SVM) with a non-linear Radial Basis Function (RBF) kernel.

TABLE II: Results for land-use sensing

	Precision	Recall	F1-score
Residential	0.69	0.78	0.73
Commercial	0.73	0.69	0.71
Manufacturing	0.86	0.81	0.82
Parks	0.77	0.69	0.72

Figure 1 and Table II gives the results of our land-use sensing experiments. In terms of classes, relatively low scores of Residential and Commercial areas could possibly be explained by the high concentration of high-rise buildings in New York City, particularly Manhattan, where the same building offers multiple spatial contexts. This also explains the high scores obtained for Manufacturing areas. However, high rise buildings do not explain why Parks do not yield better scores than Commercial and Residential areas. This is all the more

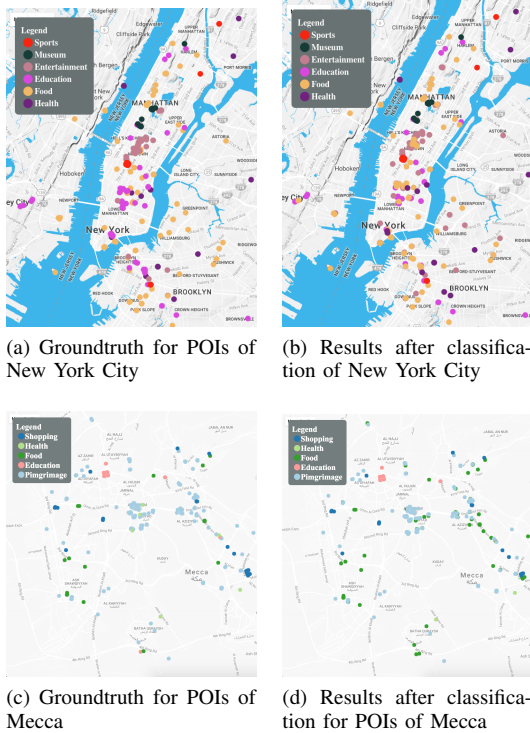


Fig. 4: Results for Points-of-Interest for New York City and Mecca are shown

surprising given that in Section III-B, Parks distinctly stand out from pockets of Residential and Commercial areas. Generally speaking, the nature of the city can have a considerable impact on the outcome of the urban sensing techniques we present here. We posit that results for urban sensing would be markedly better for more spread out cities with a lower concentration of high rise buildings.

C. Points of Interest Sensing

Finally, we used the same techniques to infer types of individual establishments in the city. Furthermore, to demonstrate that our techniques are language agnostic and work well for diverse contexts, we conducted our experiments for two highly divergent cities, with different dominant languages, in different parts of the world: New York City, US where most tweets were in English and Mecca, Saudi Arabia where most tweets were in Arabic.

Points of Interest data – We collected points-of-interests for New York City and Mecca from Open Street Map. Points of interest offering similar contexts were grouped together into broader categories. The POI categories for New York City are: Food, Sports, Health, Education, Museum, Entertainment. For Mecca, Saudi Arabia, POI categories are: Food, Health, Education, Shopping and Pilgrimage.

Having collected the points-of-interest, we used the Space-as-Documents model to create representative text documents for each establishment. Document representation, preprocessing steps and classifier used here are same as detailed in III-B.

TABLE III: Results for POIs sensing for New York City

	Count	Precision	Recall	F1-score
Museum	79	0.91	0.9	0.91
Health	85	0.81	0.88	0.85
Sports	84	0.89	0.86	0.87
Entertainment	282	0.79	0.58	0.67
Education	326	0.62	0.69	0.65
Food	682	0.53	0.62	0.57

TABLE IV: Results for POIs sensing for Mecca, Saudi Arabia

	Count	Precision	Recall	F1-score
Food	26	0.90	0.96	0.93
Health	15	0.89	0.93	0.91
Education	13	0.86	0.88	0.87
Shopping	26	0.77	0.73	0.75
Pilgrimage	150	0.73	0.66	0.69

Figure 4, Table III and Table IV give results for our experiments. Differences in scores for same kinds of spatial contexts can be attributed to the unique properties of cities in terms of their spread, structure and density. Despite the same models and techniques, note how overall the scores for points of interest sensing are higher as compared to scores for land-use sensing. We postulate that the more specific a spatial context, the more predictable the behavior observed. Regardless, the fact that the same techniques are effective at urban sensing for two radically divergent cities, offering widely different spatial contexts, and data in two different languages, is testament to the wide-applicability and effectiveness of our models and techniques.

IV. CONCLUSION

In this paper, we solve problems of urban sensing using the intuition that context influences behavior. The results of our experiments validate our intuitions and techniques. We successfully demonstrate that physical spatial context does in fact influence behavior and by studying conversations, as captured from Twitter, key insights can be gained into the spatial contexts in which they take place.

REFERENCES

- [1] *World Urbanization Prospects: The 2014 Revision ST/ESA/SER.A/366*. United Nations, Department of Economic and Social Affairs, Population Division, 2015.
- [2] J. R. Firth, *Papers in Linguistics 19341951*. Oxford University Press, 1957.
- [3] M. M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008. [Online]. Available: <http://dx.doi.org/10.1109/MPRV.2008.80>
- [4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *27th Annual Conference on Neural Information Processing Systems 2013.*, 2013, pp. 3111–3119.
- [5] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, 2014, pp. 1188–1196.
- [6] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000. [Online]. Available: <http://dx.doi.org/10.1109/34.868688>
- [7] N. Y. The Department of City Planning, “Open data,” <http://www1.nyc.gov/site/planning/data-maps/open-data.page>, 2016, accessed: 2016-09-01.