Exploring the geo-temporal patterns of the Twitter messages

Muhammad Adnan[*], Guy Lansley[†], Paul A. Longley[‡]

Department of Geography, University College London, Gower Street, London, WC1E 6BT.
Email: m.adnan@ucl.ac.uk, g.lansley@ucl.ac.uk, p.longley@ucl.ac.uk

November 07, 2015

Summary

This paper explores the data recorded through the Twitter social media service. In particular we are interested in the analysis of the content of Tweet messages. A large corpus of Twitter messages was analyzed and Index of Dissimilarity measure was used to identify interesting words having spatial concentrations. The paper presents an initial exploration of the spatial and temporal pattern of the identified interesting words. At the finest geographical level, this type of analysis can gage very useful information to local planners in general and retail planners in particular.

KEYWORDS: Social Media, Geo-Temporal Analysis, Twitter, Content Analysis

1. Introduction

Recent years have seen an increased use of social media data as a cheaper alternative to more traditional methods of market research. Social media services generate a large quantity of data every day and some of the data is available through their Application Programming Interfaces (APIs). Social media services such as Twitter allow users to share information via short messages. These services are used not only for communicating with friends, family, and colleagues, but also for real-time news feeds and content sharing about venues (Pennacchiotti and Popescu, 2011). According to recent figures, the Twitter service has more than 200 million active users around the world (Twitter, 2012a). Its major user base is in European countries: in the context of the present paper, usage in the city of London, New York and Paris is the 3rd, 5th, and 7th highest in the world (Bennett, 2012). Twitter users generate a huge quantity of data every day, and our motivation here is to explore the geo-temporal patterns which exist in the text messages themselves. This paper presents an analysis of a large dataset of Twitter messages by the identification of a range of interesting words. Words were assigned to different categories and an initial exploration of the spatial and temporal pattern of the categories is presented. At the finest geographical level, this type of analysis can provide very useful information to local and retail planners.

Analysis of the social media content is a promising research area. Whilst past research on the Tweets' content has emphasized on exploring the sentiments users express in their messages, there has been limited attempts to link the geography of user generated topics across space to land use and activity. Some related work includes: the use of social media messages to classify areas into homogeneous groups (Birkin et el, 2013), the analysis of the personal information included in the tweet messages (Humphreys et el, 2013), historicizing Twitter within a longer historical framework of diaries

[*] m.adnan@ucl.ac.uk
[†] g.lansley@ucl.ac.uk
[‡] p.longley@ucl.ac.uk

(Humphreys et el, 2014), the content analysis of Tobacco-related Twitter posts (Myslín et el, 2012), and a forecasting model to predict the spread of a news (Naveed et el, 2011).

This paper is comprised of 5 sections. Section 2 of this paper describes the data used in the analysis. Data processing is described in the section 3, while section 4 and 5 present the results and conclusion.

2. Data

The Twitter Streaming API (Twitter, 2012b) can be used to download a 1% sample of the geotagged tweets. For this paper, the Twitter Streaming API was used to download geo-tagged Tweets for the Greater London during July to December, 2013. The fields downloaded from the API included the user name, latitude and longitude from which the Tweet was sent, time and tweet message content. A total of 4.6 million (4,633,139) geo-tagged Tweets were downloaded. These tweets were sent by a total of 272,248 unique users. Following map (Figure 1) shows a map of the 4.6 million tweets. This map shows that more Tweets were sent by users located in the central part of the city than the surrounding areas of Outer London.

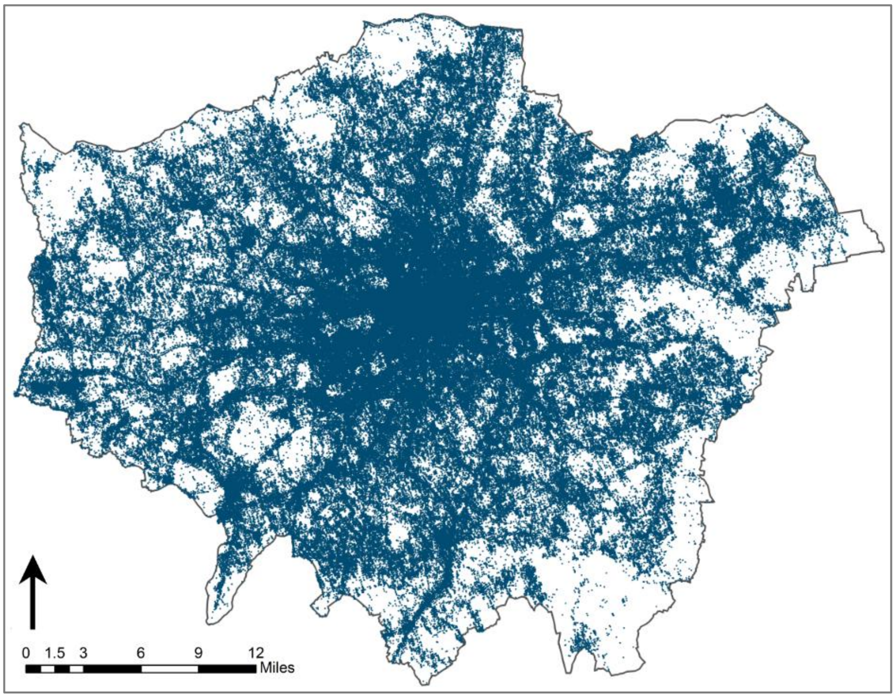


Figure 1: The Greater London geography of the 4.6 million tweets

Few users sent more tweets than others. 2,000 or more tweets were sent by the top 45 users. Following figure (2) shows the number of tweets by individual users.

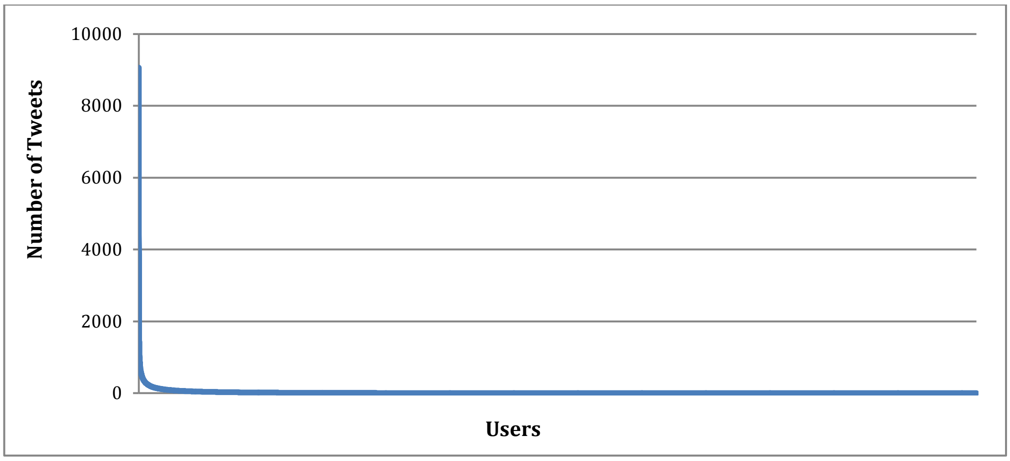


Figure 2: Number of tweets by individual users

3. Data processing

In the first step, 4.6 million tweets messages were divided into a series of ‘words’ i.e. a group of characters separated by a full-stop, comma, semi-colon, colon, apostrophe, or double quotes. This resulted in a dataset of 35,028,273 words. For the investigation of the spatial patterns of individual words, all the words were aggregated to 633 wards in the Greater London. For each word (*y*), an Index of Dissimilarity (Birkin et el, 2013) was calculated across the 633 wards. The Index of Dissimilarity is defined in the following equation.

$$\theta\ (x,z) = 0.5\ \times \sum_{x} \left| \frac{X_x^y}{X_*^y} - \frac{X_x^*}{X_*^*} \right|$$

Where x = *(1,…..,633)* wards in the Greater London and an asterisk (*) denotes summation across a missing index. The resulting Index of Dissimilarity value for each word (*y*) is a standardized value between 0 and 1. Where 0 indicates a uniform distribution and a 1 indicates a spatial concentration.

Index of Dissimilarity was calculated for each of the 35,028,273 words in the dataset. In the second step, in order to select the words which are spatially concentrated, the words having Index of Dissimilarity less than 0.5 were deleted from the database. This resulted in 122 remaining words which are listed in the following table (1). The table also assigns each word to one of the 8 distinct categories.

Table 1: 122 spatial concentrated words

Categories	Words
Travel	LHR, PANCRAS, PADDINGTON, HEATHROW, RAILWAY, UNDERGROUND, FLIGHT, STATION, @HEATHROWAIRPORT, AIRPORT, TERMINAL, TUBE
Sports	#THFC, FULHAN, #ARSENAL, #LFC, #AFC, #ASHES, #CFC, @ARSENAL, CHELSEA, SPURS, FOOTBALL, #MUFC
Places in London	HOUSNLOW, MARYLEBONE, MIDDLESEX, BROMLEY, GREENWICH, ISLINGTON, SHOREDITCH, OXFORD, PICCADILLY, WHARF, KINGSTON, SHARD, HACKNEY, BRIXTON, BRICK, MARKET, KENSIGNTON, LEICESTER, KNIGHTSBRIDGE, CROYDON, HAMMERSMITH, CIRCUS, TOTTENHAM, WATERLOO, NOTTING, COVENT, REGENT, ARENA, WESTFIELD, ROMFORD, CAMDEN, RICHMOND, CLAPHAM, STRATFORD
Tourism	MUSEUM, TOWER, GALLERY, BRIDGE, PALACE, ROYAL, HOTEL, COURT, TRAFALGAR, HYDE, WESTMINSTER, ALBERT, BUCKINGHAM
Food & Drink	@STARBUCKSUK, STARBUCKS, COCKTAILS, BAR, COSTA, PUB, DRINK, COFFEE, JUICE, CAFE, MCDONALDS, COOKING, RESTAURANT

Leisure	LOUNGE, STUDIOS, THEATRE, PARK, EVENT, CINEMA, XFACTOR, KITCHEN, HOLIDAY, XBOX, HANGING, GARDEN, SHOPPING, MUSIC
Emotions	ENJOYED, #EXCITED, OMG, MISSING, SURPRISED, DISGUSTING, EMBARRASING, ANNOYING, GAY, MADNESS, WTF, FANTASTIC, SHOCKING, RIDICULOUS, BORED, AWFUL, HAPPINESS, PLZ
Other	GOODNIGHT, DUDE, DAD, DADDY, BOYS, FAMILY, FRIEND

4. Results and Discussion

Following figure (3) shows an example of the spatial concentration of the words. This figures shows two maps of the individual tweets where ‘TRAFALGAR’ (map on the left) and ‘LHR’ (map on the right) were mentioned in the tweet messages. The Index of Dissimilarity value for both the words was 0.833 and 0.96 respectively, indicating a spatial concentration of the tweets. This could also be seen in the maps.

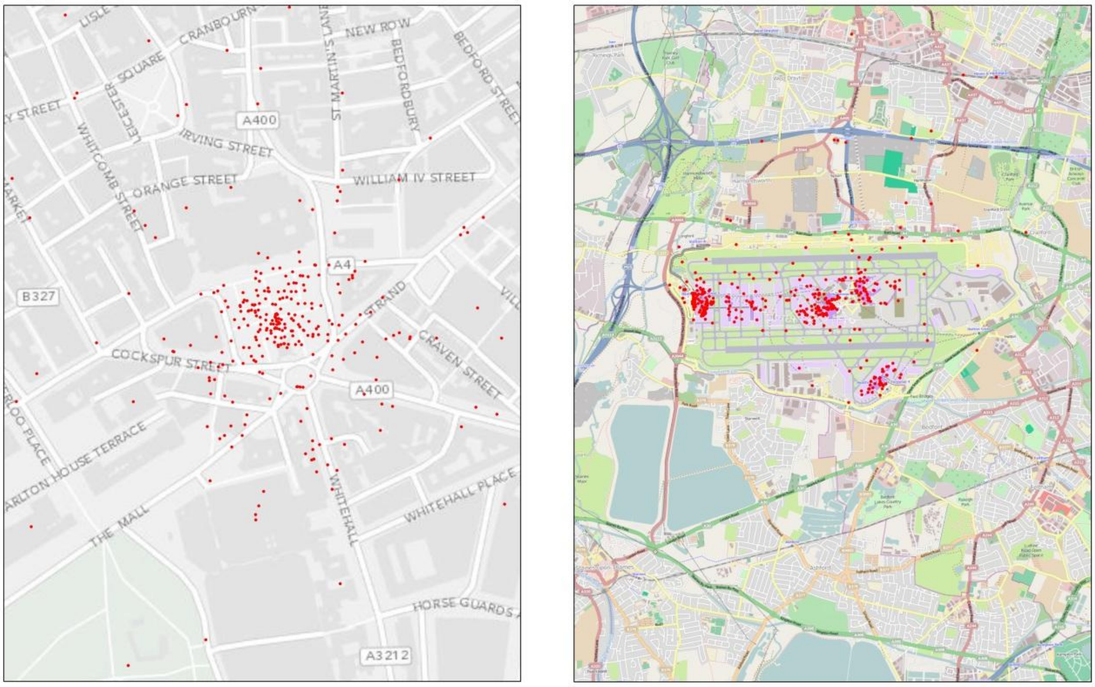
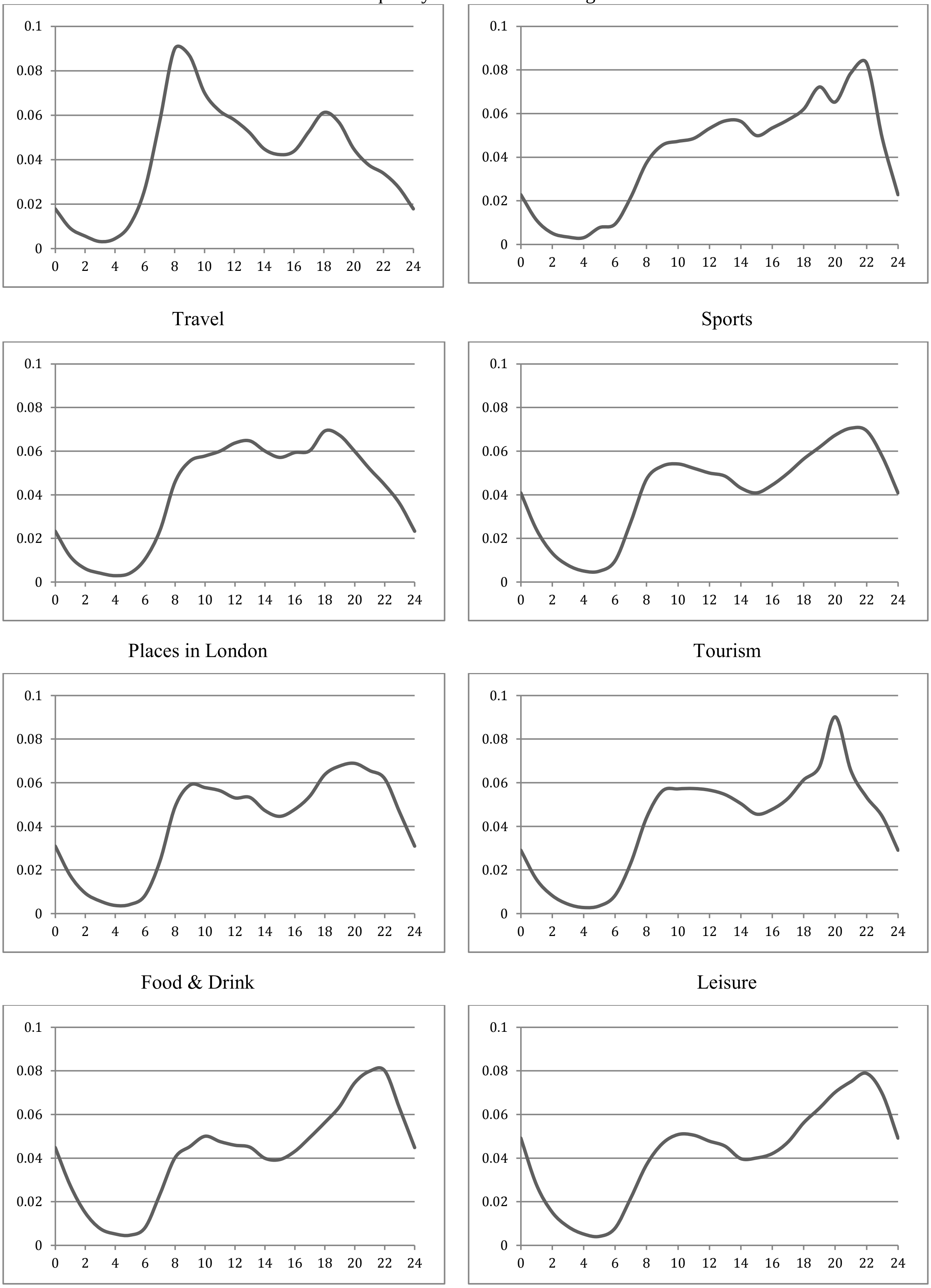


Figure 3: Tweets around the area of Trafalgar Square (left) and London Heathrow Airport (right)

The following table (2) shows the temporal graphs of the 8 word categories listed in section 3. The temporal graphs show the distinct temporal patterns of these categories. Words of the ‘Travel’, ‘Sports’, and ‘Leisure’ categories have the most distinct patterns. There is high number of tweets mentioning ‘Travel’ category words during the morning and evening rush hours. More tweets of the ‘Sports’ and ‘Leisure’ category words are sent during the night time. There are also more tweet mentions of the tourist places after 3pm during the day.

Table 2: Temporal graphs of the word categories. X-axis represents the hours of the day and Y-axis represents the frequency of the tweet messages



Travel

Sports

Places in London

Tourism

Food & Drink

Leisure

Emotions

Other

These temporal graphs show the footprints of the Twitter activity throughout the city. These also show an overall pattern of the behavior of the users in the Greater London.

5. Conclusion and future work

This paper has presented a preliminary analysis of the Twitter messages to explore the inherent spatial and temporal patterns of activity. A large dataset of Twitter messages was analyzed and decomposed into 35,028,273 words. For each word, the Index of Dissimilarity was calculated to identity interesting words having spatial concentrations. This resulted in a total of 122 words which were assigned to 8 distinct categories. The paper has also presented an initial exploration of the spatial and temporal pattern of the word categories.

This is a very promising research area, and we plan to enhance this work in the future. We plan to perform a fine scale temporal activity pattern analysis on the dataset to identity the areas of distinct attributes and behaviors e.g. the areas of leisure activities vs. work place areas. We also plan to use various topic unsupervised modelling techniques such as Latent Dirichlet Allocation (LDA) to generate topics in small geographical areas, and analysing the temporal variations in topic formulation and popularity, both daily temporally and seasonally.

Acknowledgements

This work was completed as part of the ESRC research Grant "Retail Business Datasafe" (ES/L011840/1).

References

Bennet, S. 2012. Revealed: The Top 20 Countries and Cities of Twitter [STATS]. Retrieved 31st December, 2012, from http://www.mediabistro.com/alltwitter/twitter-top-countries_b26726.

Birkin, M., Harland, K., Malleson, N. (2013). The classification of space-time behavior patterns in a British city from crowd-sourced data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7974, pp.179-192.

Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2013. Historicizing New Media: A Content Analysis of Twitter. *Journal of Communication*, 63, 413-431.

Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2014. Twitter: a content analysis of personal information. *Information, Communication & Society*. 17 (7).

Myslín, M., Zhu, Shu-Hong., Conway, Michael. 2012. Content Analysis of Tobacco-related Twitter Posts. In the proceedings of the 2012 International Society for Disease Surveillance Conference.

Pennacchiotti, M. and Popescu, A. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the Fifth International AAAI conference on Weblogs and Social Media.*

Naveed, N., Gottron, T., Kunegis, Jérôme., Alhadi, Arifah Che. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In the proceedings of the WebSci’11. Koblenz, Germany. June 14-17, 2011.

Twitter. 2012a. What is Twitter ?. Retrieved 31st December, 2012, from https://business.twitter.com/basics/what-is-twitter/.

Twitter. 2012b. The Streaming APIs ?. Retrieved 22nd January, 2012, from https://dev.twitter.com/docs/streaming-apis.

Biographies

Muhammad Adnan is a Senior Research Associate at Consumer Data Research Centre, University College London. His research interests are in data mining, social media analysis, and visualisation of large spatio-temporal databases.

Guy Lansley is a Research Associate at the Consumer Data Research Centre, UCL, an ESRC Data Investment. His previous research has included exploring the temporal geo-demographics derived from social media data, and identifying socio-spatial patterns in car model ownership in conjunction with the Department for Transport. Whilst, his current work entails exploring population data derived from large consumer datasets.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal Environment and Planning B and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.