Spatio-Temporal Patterns of Passengers' Interests at London Tube Stations

Juntao Lai[*1], Tao Cheng[†1], Guy Lansley[‡2]

[1] SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental &Geomatic Engineering, University College London
[2]Department of Geography, University College London

April 15, 2015

Summary

With as many as 3.5 million passengers using the London underground system every day, it is desirable to examine and understand their interests and opinions, and to harness this information to improve the services of Transport for London (TfL). This research aims to achieve a better understanding of passengers' interests by harvesting text from geo-tagged Tweets sourced over a four week period in 2014 from the vicinity of the stations. An unsupervised topic modelling method Latent Dirichlet Allocation (LDA) is used to generate topics, and k-means is used to cluster stations in order to understand the overall patterns of topics.

KEYWORDS: social media data, spatial-temporal computation, K-means, topic modelling

1. Introduction

More than 3.5 million passengers use the London underground system every day, and more than 1 billion journeys are made every year (TfL, 2014). It is desirable to examine and understand the interests of passengers, and harness this information to improve the services of Transport for London (TfL), or make fuller use of the commercial advertisement potential inside the underground stations. However, it is difficult to collate such information directly from the millions of underground users. With the recent advance of smartphones and internet coverage, microblogging applications such as Twitter have been widely used. Twitter is real-time and is more widely spread compared with other blogging systems and traditional media (Java et al., 2007; Zhao et al., 2011). This creates opportunities to understand the interests of populations in space and time.

Based on the assumption that Tweets around underground stations are likely to be posted by underground users, this research aims to achieve a better understanding of the interests of London Underground passengers. This is achieved by harvesting text from geo-tagged Tweets sourced over a four week period in 2014 from the vicinity of the stations. An unsupervised topic modelling method Latent Dirichlet Allocation (LDA) is used to generate topics from Tweets, and k-means is used to cluster stations in order to understand the overall patterns of topics.

2. Spatio-Temporal patterns of Twitter topics at London tube stations

Here we describe the major steps to detect the topics of tweets and present their spatial-temporal distribution. The results are further clustered in order to understand the overall patterns of topics.

[*] Juntao.lai.13@ucl.ac.uk
[†] Tao.cheng@ucl.ac.uk
[‡] G.lansley@ucl.ac.uk

2.1. Data Description

The Twitter data is downloaded from the Twitter Streaming Application Programming Interface (API) service (Twitter Developers, 2012). The temporal range of the data is 4 weeks, from Feb 25 to Mar 24 in 2014. The data has been divided into two groups: the weekday group (Tuesdays, Wednesdays and Thursdays) which has 887,600 Tweets; and the weekend group (Saturdays and Sundays) which has 687,700 Tweets.

2.2. Assign Tweets to unique tube stations

Given the density of London tube stations, a strategy is required to assign individual Tweets to unique tube stations. This is achieved by defining a unique coverage for each tube station, which is generated by using Thiessen (Voronoi) Polygons (Franz, 1991). Tweets are assigned to the station corresponding to the Thiessen polygon they fall within. Since we are mainly interested in topics discussed near or at the tube station, the Tweets within a 10-minute walking distance are extracted as the Tweets for further analysis.

2.3. Generate hot topics from Twitter data using LDA

A script using the R package named "tm" (Feinerer, 2014) was used to remove the "noise" from the Twitter data, which includes the process of removing the whitespaces, numbers, punctuations and stopwords, also converting all the upper case to lower case. The process of "stemming", which required an R package named "SnowBallC" (Bouchet-Valat, 2014), was also used to reduce inflected words to their stem form, by removing suffixes of the words.

After text cleaning, a topic modelling method named Latent Dirichlet allocation (LDA) is used to generate topics. LDA is an unsupervised generative model which can be used to classify text documents (Blei et al., 2003). A document is made up of groups of words which may belong to different topics. A topic is a bag of words, with corresponding probabilities of each word belonging to this group. LDA is a process that tries to backtrack from the documents to find a set of topics that are likely to have been generated by the collection. LDA represents documents as mixtures of topics that spit out words with certain probabilities (Chen, 2011). The main idea of LDA topic modelling is that the words that appear together many times in the documents are assumed to be related or present similar meaning, and are therefore more likely to be assigned to the same topic. It is more efficient and objective than manually classifying the text.

LDA has been used widely for news and text analysis, but its application on analysing short and informal documents like Tweets has only been implemented recently for detecting and analysing big events, such as earthquakes (Caragea, et al., 2011), the outbreak of flu (Chew et al., 2010), or special events (Cheng and Wicks, 2014). Here we utilise techniques of semantic analysis of Tweets to investigate the daily interests (topics) of underground users. An R package named "lda" (Chang, 2013), written by Jonathan Chang, was used to generate the topics from Tweets in this study.

2.4. Clustering stations using K-means

A general spatial temporal pattern of the topics can be revealed by focusing on the most frequent topics, but this only represents one variable among all informative ones (topics). In order to investigate the interest of the users more comprehensively, the influence of all topics that have been generated should be included. Therefore, the stations having similar topic distributions need to be classified into groups.

The classification of the stations based on their topic distributions was carried out using K-means clustering method. K-means clustering (MacQueen, 1967) enables the data to be clustered into a pre-specified number of groups. Based on the results generated by LDA, a table was created to store the topic distributions in both spatial and temporal dimensions. The percentage of the Tweets belonging to each topic were input as variables.

3. Results

The Twitter data was clipped by station assigned polygons and each Tweet was joined with one unique station name as an attribute. After data cleaning and formatting, the Tweets were fitted into the LDA topic model, then the words were assigned into groups by the model. After interpreting the groups using their top words, 10 meaningful topics were selected and labeled manually.

Using the topic distributions calculated from LDA, the stations are classified into groups. The categorization of the stations after clustering are displayed on the maps (Figure 1), it can be seen that the major groups of the users on weekends are cluster 1 and cluster 3. According to the cluster center information, cluster 1 has great value in "sports", hence it could be interpreted as sports fans. Cluster 3 could be represented as working population due to its high proportions in "traffic" and "work-home lifestyle". The other big group is cluster 5. Based on its big numbers in "tourism & travel" and "food & drinks", and relatively low values in other topics, this group is more likely to represent tourists. The distributions of clusters on Figure 1 reflect reasonable patterns of those groups. For example, the central areas are mainly occupied by tourist groups, and people are more likely to talk about sports in the afternoon than in the morning.

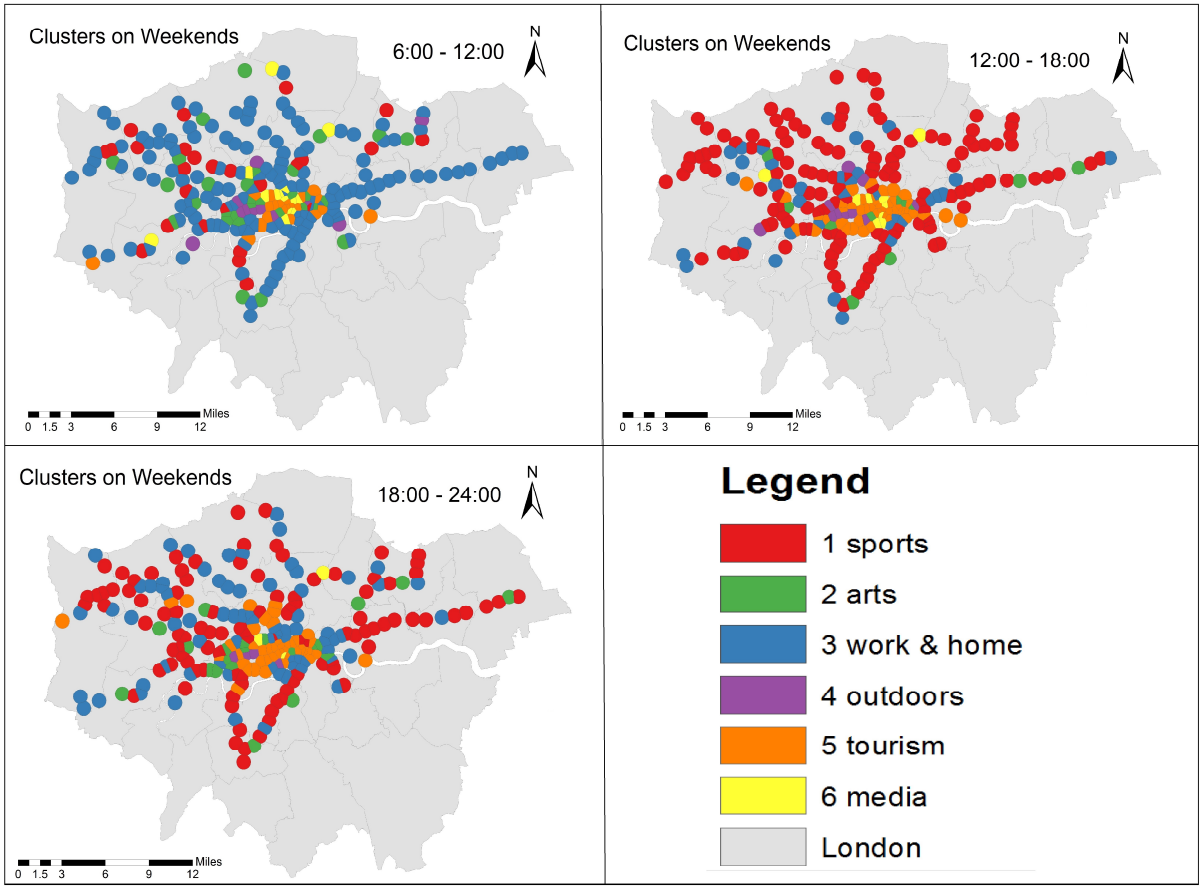

Figure 1: Clusters in Different Time Periods on Weekends

4. Conclusion

This study demonstrated the effectiveness of LDA topic modeling in extracting topics from Twitter data, as previously hypothesized. However, many extensions to LDA were suggested to improve the quality of extracting and labelling the topics, such as supervised LDA (McAuliffe, 2007) and Labeled-LDA (Ramage, 2009) which may fit Twitter data better.

This paper also presented as a case study that successfully generated the main topics that underground

passengers discuss using Twitter data, and then explored the distributions of these topics spatially and temporally. Furthermore, these topics are reasonable and distinguishable in terms of representing daily activities of the public. Finally, the stations were classified into several groups in each time periods according to their topic distributions, which could be helpful to better understand the interest of underground passengers from a statistical perspective.

5. Biography

Juntao Lai is a PhD student in department of Civil, Environmental and Geomatic Engineering at University College London. His research interest includes semantic and sentiment analysis of social media data and spatial-temporal analytics. His current work is investigating the impact of media to the public satisfaction on policing using Twitter data.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab for Big Data Analytics (http://www.ucl.ac.uk/spacetimelab), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

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.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003) "Latent dirichlet allocation." *the Journal of Machine Learning Research* 3: 993-1022.

Bouchet-Valat, Milan. (2014) *"Package "SnowballC". R package version 0.5.1.* [ONLINE] Available at: http://cran.r-project.org/web/packages/SnowballC/ [Accessed 10 August].

Caragea, Cornelia, et al. (2011) "Classifying text messages for the Haiti earthquake." *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*.

Chang, Jonathan. (2013) *"Package 'lda'". R package version 1.3.2.* [ONLINE] Available at: http://cran.r-project.org/web/packages/lda/lda.pdf [Accessed 16 July 14].

Chen, Edwin. (2011) *Introduction to Latent Dirichlet Allocation.* [ONLINE] Available at: http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/ [Accessed 15 July 14]

Cheng, T. & Wicks, T. (2014). Event Detection using Twitter: A Spatio-Temporal Approach. Plos One, 9(6), e97807

Chew, Cynthia, and Gunther Eysenbach. (2010) "Pandemics in the age of Twitter: content analysis of . Tweets during the 2009 H1N1 outbreak." *PloS one 5.11: e14118*.

Feinerer, Ingo. (2014) *"Introduction to the tm Package Text Mining in R." Comprehensive R Archive Network.* [ONLINE] Available at: http://cran.r-project.org/web/packages/tm/index.html [Accessed 16 July].

Franz Aurenhammer (1991). Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys, 23(3):345–405*

Java, Akshay, et al. (2007) "Why we twitter: understanding microblogging usage and communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.* ACM.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).*

Mcauliffe, Jon D., and David M. Blei. (2008) "Supervised topic models." *Advances in neural information processing systems (pp. 121-128).*

Ramage, Daniel, et al. (2009) "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics.

Transport for London. (2014). *London Underground – Factsheet.* [ONLINE] Available at: http://www.tfl.gov.uk/cdn/static/cms/documents/lu-factsheet-jan2012.pdf [Accessed 29 June 14].

Twitter Developers. 2012. *The Streaming APIs.* [ONLINE] Available at: https://dev.Twitter.com/docs/streaming-apis [Access 03 July 14]

Zhao, Wayne Xin, et al. (2011) "Comparing Twitter and traditional media using topic models." *Advances in Information Retrieval. Springer Berlin Heidelberg. 338-349.*