Spatial Narrative Construction using Thematic KDE

N. C. Bennett[1], D. E. Millard[2], D. J. Martin[1], and P. Amirian[3]

[1](n.bennett,d.j.martin)@soton.ac.uk

[2]dem@ecs.soton.ac.uk

[3]pouria.amirian@os.uk

*Email: n.bennett@soton.ac.uk

Abstract

The production of large volumes of social media Volunteered Geographic Information (VGI) provide researchers with a large-scale data source for understanding social activities and behaviours. Previous investigations focus on extracting singular words to describe activities and locations, removing vital contextual and narrative information. This work is a narrative-centric approach, using subtext analyses to discover the inherent location-based themes present within social media discourse on Twitter. This work presents a paradigm shift away from singular to rich, narrative-centric descriptors by proposing a new methodological framework and applying it to a database of over two million tweets. These narratives can subsequently be used to bring greater understanding to how people interact with their environment.

1 Introduction

Social media platforms allow users to reflect upon their surroundings, often accompanied by a location tag. A popular example of this type of platform is Twitter. As of June 2016[1], Twitter is used monthly by 313M users (both human and automated) with 82% of these users tweeting from mobile devices. With the ever increasing saturation of smartphone devices, this represents a large proportion of users able to produce geotagged information. This offers an unprecedented insight into spatio-temporal activities in real-time (Johnson et al., 2016). Previous research focused on aggregating tweets into topics but frequently condensed them into one-word summaries either for visualisation or as a product of the analysis at the expense of contextual information (Steiger et al., 2015; Adams and Remiro-Azócar, 2016); however, it is this very information that can provide rich narrative understanding, providing both academic and commercial researchers with dynamic social information. Twitter is a useful resource for obtaining spatio-temporal data; its Streaming API collects between 1-2% of raw tweets, but research has shown it collects over 90% of all georeferenced tweets (Morstatter et al., 2013).

This work adapts the view of Russian formalist Tomashevsky (1965) which states narratives in literature are comprised of themes, motifs and features, and applies it to modern social media posts (tweets) from Southampton and wider Hampshire, UK. This novel approach focuses on words (features) within tweets that relate to a high-level conceptual theme and represents their spatial

[1]Date and details taken from http://about.twitter.com/company [Accessed 01-02-2017]

components. To do this, a novel methodology is proposed that constructs spatial narratives based on the subtext (the themes) of geolocated tweets.

2 Related Work

Mobility pattern analysis is a key research field from which this work draws. Work into analysing spatio-temporal mobile phone data provides temporally accurate conclusions on the whereabouts of users before, during and after big events (Mazimpaka and Timpf, 2017). However, with these studies the spatial element is imprecise and the semantic information is completely lacking. Attempts at using social media data to supplement mobile phone data, in this case analysing geolocated tweets, have allowed researchers to understand environmental factors that impact mobility (Wu et al., 2015) and locate catchment areas for shopping centres (Lloyd and Cheshire, 2017) by matching a secondary dataset of location-identifying words to the tweets.

Research applying topic analysis to social media content has shown how a set of one-word summaries can be produced. Topic modelling algorithms such as latent Dirichlet allocation (LDA), outlined in Blei et al. (2003), offer an aggregated overview of prevalent topics within the dataset using unsupervised machine learning. LDA has been used to classify work- and home-related tweets and compare them with expected locations derived from census data (Steiger et al., 2015), to compare the relative importance of local and national news (Bennett et al., 2016) and to form the basis of event detection methodologies (Weng et al., 2011; Atefeh and Khreich, 2015). However, aggregation removes key narrative information by reducing the vast contextual information to a one-word summary. To represent the data without aggregation, this work maps multiple features onto appropriate high level themes, resulting in a more continuous thematic view, rather than choosing the most popular topic.

Due to the limitations of topic modelling methods, this work proposes term expansion as an improvement upon existing semantic enrichment methods. Previous work on term expansion focuses on queries, for both textual and image results, attempting to create complex ontologies and word graphs for linguistic expansion of query terms (Gong and Liu, 2009; Martins et al., 2013). Such complex graphs are not needed for spatial narratives, instead we have used a more specific structure that captures high level themes, the motifs that connote them, and the features that denote those motifs (Hargood et al., 2010).

3 Proposed Methodology

The proposed novel methodology, outlined in Figure 1, is derived from previous work on mobility analysis and topic modelling but differs in several key ways. Firstly, mobility pattern studies focus on spatial and temporal analysis to the detriment of semantic information. Whilst these studies often only have point data, such as cellular or public transport locations (Bagrow et al., 2011), previous studies concerned with mobility patterns from social media often reduce or remove textual information. Studies that do include semantic information (Steiger et al., 2015; Lansley and Longley, 2016) often aggregate text to the detriment of narrative context.

This methodology differs from the above studies as it negates the need for aggregation by tagging each tweet with its associated themes, therefore capturing the essence of the tweet rather than highlighting specific words or phrases.

3.1 Pre-Processing

The tweets were collected using a boundary box search over Hampshire, UK, with Southampton as its centre. As this paper focuses on extracting narrative content, obtaining a database with precise and authentic tweets is necessary, requiring the removal of irrelevant (spam) tweets. To discover these tweets, semantic keywords[2] are used as an initial indicator, with tweets containing them being discarded. If a user has more than ten matching tweets, the account is removed from the dataset.

3.2 Thematic Tagging

A challenge for natural language processing is obtaining an objective semantic dataset with which the researcher can compare their collected data. Approaches to manually assigning and labelling data have proven to be highly subjective (Habernal et al., 2013). Therefore, the preferred method is to objectively acquire a dataset of thematic words. However, some elements of subjectivity such as choosing a set number of themes and reviewing the attribution process is necessary.

To create the thematic database a theme, such as 'commerce', is entered into an online thesaurus[3]. The returned results are collected and the more strongly associated synonyms are themselves searched for and their top results collected. This creates a database of associated words from which those that only appear once are discarded. The remaining words are then assigned to the 'commerce' theme. The tweets are searched for matching content and positive results are tagged with the relevant theme. This is proposed as an appropriate method as it is predominantly objective. This also circumvents the previously discussed issues over tokenising and aggregating tweets. Furthermore, previous work often used WordNet[4] which offers holonyms and hypernyms as subsets of synsets (their term for a set of synonyms) but are much more complex than is necessary for this task.

3.3 Thematic KDE

From the tagged database, those tweets that have coordinates in their metadata are parsed into a kernel density estimation (KDE) algorithm to map hotspots of activity. Further to straightforward spatial mapping, the allocation of thematic tags allows for subsequent thematic mapping. KDE smooths erratic data and removes noise inherent in geolocation data, forming a probabilistic clustering of popular tweet locations. A key disadvantage of KDE is the necessity of a bandwidth metric; this metric determines the extent to which the data are smoothed. Over-smoothing can dilute smaller hotspots and under-smoothing can merge individual hotspots within a small area into one large, amorphic cluster; both methods subsequently skew the analysis. Therefore, repeat experiments with a range of values are required to select the optimum bandwidth.

[2]Outlined in https://blog.hubspot.com/blog/tabid/6307/bid/30684/The-Ultimate-List-of-Email-SPAM-Trigger-Words.aspx [Accessed 01-02-2017]

[3]Available at http://www.thesaurus.com/

[4]http://wordnet.princeton.edu [Accessed 01-02-2017]

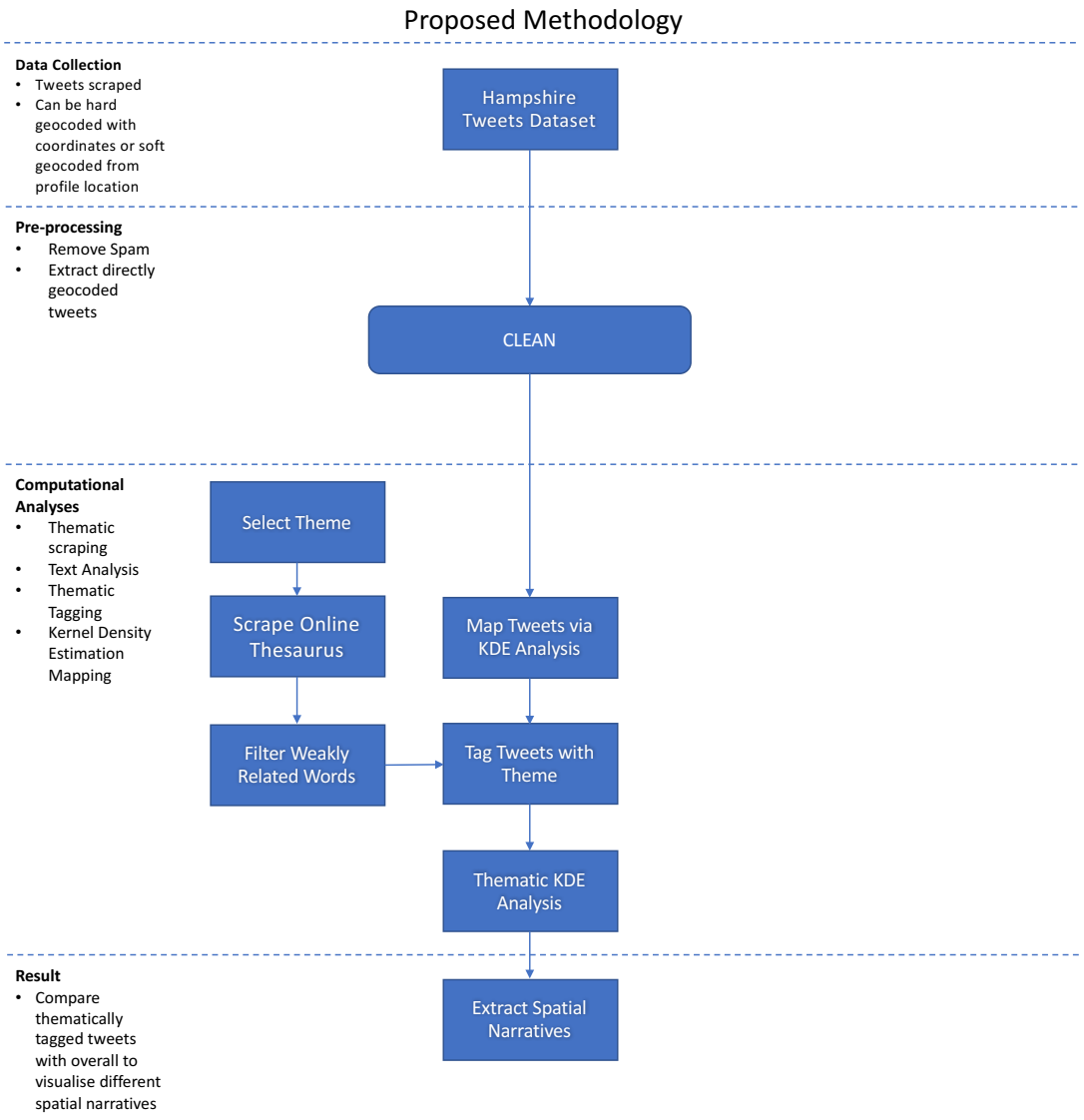


Figure 1: Flowchart describing the methodological approach.

4 Results & Discussion

This section outlines the results of the thematic kernel density estimation analysis. All maps were created in ArcGIS via ArcMap[5].

Tweets	Raw	Cleaned
Total	1,988,062	1,849,323
Geotagged	17,897	12,811

Table 1: Table showing tweet numbers before and after cleaning.

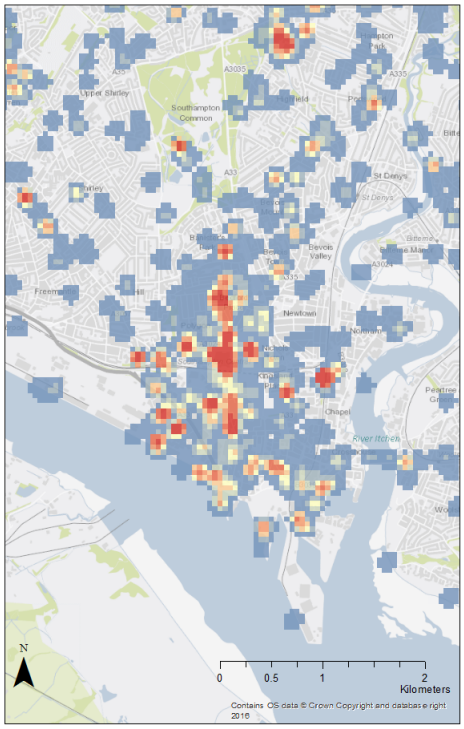
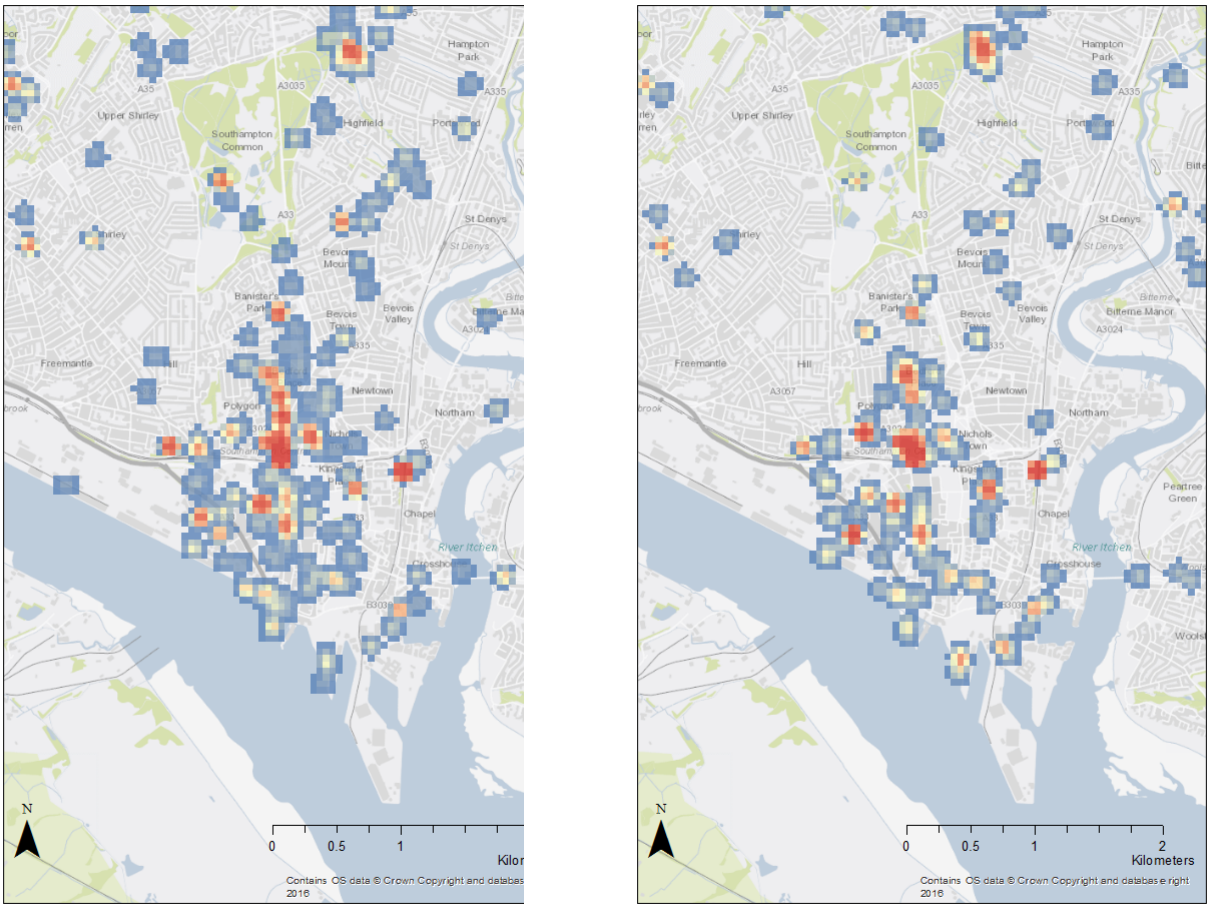


Figure 2: Cleaned KDE heatmap of 12,811 geotagged tweets, centred on Southampton, UK.

[5]http://desktop.arcgis.com/en/arcmap/



(a) Cleaned KDE heatmap of 1,844 tweets tagged with the commerce theme.

(b) Cleaned KDE heatmap of 1,018 tweets tagged with the entertainment theme.

Figure 3: Two cleaned maps placed side-by-side for comparison. Their profiles are subtly different, the subtly likely caused by the urban geography creating an social space for both themes.

It is clear that the commerce tweets shown in Figure 3a follow a similar distribution to the overall dataset shown in Figure 2, with large clusters seen over the city centre. However, when comparing the distribution seen in Figure 3b, the clusters are tighter around individual entertainment buildings, exemplified by fewer clusters and noticeable gaps along the centre. This is likely reflecting the respective patterns of population densities during commercial and leisure activities. Despite this, the north-south trend does reflect the urban geography, thus is useful for understanding city-level mobility patterns (Mislove et al., 2011).

5 Conclusions

This paper was successful in producing a thematic kernel density estimation analysis of tweets from Southampton and wider Hampshire. Whilst crawling through a thesaurus allowed for the collection of an objective dataset of semantically-linked thematic words, some of the more common words caused more tweets to be tagged than expected (for instance 'work' appeared in both commerce and entertainment but the tweet rarely applied to either). Future work will refine the tagging method to exclude common words and those that belong to several different themes. Future work will also concentrate on automating motif selection, allowing for a more conceptual rather than prescriptive thematic tagging process.

6 References

Adams, P. and Remiro-Azócar, A. (2016). *City-wide Mobility Mapping Using Social Media Communications*. PhD thesis, Unversity of Bath.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, 31(1):133–164.

Bagrow, J. P., Wang, D., and Barabási, A. L. (2011). Collective response of human populations to large-scale emergencies. *PLoS ONE*, 6(3):e17680.

Bennett, N. C., Millard, D. E., and Martin, D. J. (2016). Narrative Extraction through the Detection and Characterisation of National and Local Events. In Gartner, G. and Huang, H., editors, *Proceedings of the 13th International Conference on Location Based Services*, pages 196–200, Vienna. Vienna University of Technology.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., and Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Gong, Z. and Liu, Q. (2009). Improving keyword based web image search with visual feature distribution and term expansion. *Knowledge and Information Systems*, 21(1):113–132.

Habernal, I., Ptácek, T., and Steinberger, J. (2013). *Sentiment analysis in czech social media using supervised machine learning*. Number June.

Hargood, C., Millard, D. E., and Weal, M. J. (2010). A Semiotic Approach for the Generation of Themed Photo Narratives. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia.*, pages 19–28.

Johnson, I. L., Sengupta, S., Schöning, J., and Hecht, B. (2016). The Geography and Importance of Localness in Geotagged Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 515–526, New York, New York, USA. ACM Press.

Lansley, G. and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58:85–96.

Lloyd, A. and Cheshire, J. (2017). Deriving retail centre locations and catchments from geo-tagged Twitter data. *Computers, Environment and Urban Systems*, 61:108–118.

Martins, F., Haslhofer, B., and Magalhães, J. (2013). Query Expansion using open Web - based SKOS Vocabularies. In *ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine.*, pages 35–38. ACM.

Mazimpaka, J. and Timpf, S. (2017). How They Move Reveals What Is Happening: Understanding the Dynamics of Big Events from Human Mobility Pattern. *ISPRS International Journal of Geo-Information*, 6(1):15.

Mislove, A., Lehmann, S., Ahn, Y.-y., Onnela, J.-p., and Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Artificial Intelligence*, pages 554–557.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*, pages 400–408.

Steiger, E., Westerholt, R., Resch, B., and Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54:255–265.

Tomashevsky, B. (1965). Thematics. In *Russian Formalist Criticism: Four Essays*, page 143.

Weng, J., Yao, Y., Leonardi, E., Lee, F., and Lee, B.-s. (2011). Event Detection in Twitter. *Development*, 11(98):401–408.

Wu, F., Li, Z., Lee, W.-C., Wang, H., and Huang, Z. (2015). Semantic Annotation of Mobility Data using Social Media. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 1253–1263, New York, New York, USA. ACM Press.