Towards a Unified Narrative-Centric Spatial Clustering Model of Social Media Volunteered Geographic Information

Nick Bennett*[1], David Millard†[1], David Martin‡[2] and Pouria Amirian§[3]

[1]Electronics and Computer Science, University of Southampton
[2]Geography and Environment, University of Southampton
[3]Ordnance Survey, United Kingdom

January 13, 2017

Summary

Social narratives are formed from interactions with the environment. Discovering these narratives allow researchers to better understand society, which has application to a variety of social and governmental interests. This work theorises a paradigm shift in how researchers can extract narratives from social media posts by instead analysing the subtext present within the data. Combining analytical methods such as adaptive kernel density estimation and natural language processing with established social theory, a new approach for subtext analysis is argued. A new analytical framework is therefore necessary and is proposed in this work.

KEYWORDS: social media, VGI, narrative space, analytical framework

1 Introduction

Narratives make sense of the world (Cambria and White, 2014). In understanding narratives we are able to gain insight into the complex social interactions that occur within a person's "activity space" (Mennis et al., 2012), defined as the physical area in which they carry out their daily activities. When these activity spaces overlap this creates a dynamic, location-based social palimpsest rich with socio-technical information. It is therefore desirable for academics and commercial researchers wishing to understand the use of space to investigate this under-researched area.

A rich information source for socially generated, geolocated information is the social network Twitter[1]. The platform allows for users to post 140 character messages and, with the user's permission,

*n.bennett@soton.ac.uk
†dem@ecs.soton.ac.uk
‡d.j.martin@soton.ac.uk
§pouria.amirian@os.uk
[1]http://www.twitter.com

attach their current location as latitude and longitude coordinates. This creates a database of social information with spatio-temporal attributes, ideal for understanding the complex nature of social interactions over time. However, extracting meaningful social and spatial information, such as location-based narratives, from tweets is a challenge due to their short nature and low volume of spatially referenced metadata.

This paper presents a paradigm shift in how spatial narratives are analysed. Drawing upon the seminal work of Tomashevsky (1965), in which he argued for subtext giving structure to narrative, this paper proposes a new perspective on narrative extraction and a new analytical framework to analyse subtext within social media volunteered geographic information (VGI) posts from Twitter.

2 Related Work

2.1 Defining Narratives

From Tomashevsky (1965)'s work on thematics, narratives are defined as comprising of features, motifs and themes, each forming the next. Narrative creation is an important component of societal and personal identity; it allows us to comprehend our surroundings and is an inherently community-building experience (Cambria and White, 2014; Farrow et al., 2015; Tamburrini et al., 2015). Narratives give structure to the complex interactions between people and place, attributing emotion and memory to particular areas. To extract these properties from social media, the computational approach of natural language processing (NLP) is required due to the volume of data produced. Through using content analysis tools nouns, adjectives, adverbs (features) such as "market" can be extracted to form motifs such as "exchanging of goods", which subsequently would feed into a theme of commerce. These themes are component parts within narratives. The subject of overlapping narratives, to which this work contributes, is a crucial area for advancing narratology.

2.2 Space and Place

As narratives are comprised of social interactions within potentially overlapping activity spaces it is thus important to understand these spaces. Activity space is the intersection of social activity and geographic location, itself comprised of space and place. The arguments for defining space and place are wide-ranging and without a concrete conclusion (Agnew, 2011; Gao et al., 2014).*Space* is generally consented to be an abstract, three-dimensional area within which concrete *places* stand (for an in-depth discussion, see Agnew (2011)).

With that in mind, a person's "activity space" (Mennis et al., 2012) can therefore be understood as an abstract area populated by intermediary spatio-temporal places. However, volunteered geographic information (VGI) such as tweets produced by humans and automated scripts only provide intermediary points from which an activity space can be inferred rather than offered. Natural language processing (NLP) can aid in extracting contextual data; for example, NLP is frequently used to extract popular threads of conversation (Hirschberg and Manning, 2015; Gu et al., 2016) and is suggested as useful in understanding people's motivations (Lloyd and Cheshire, 2017), despite the difficulties in using short and often poorly-worded texts (Maynard and Hare, 2015). Therefore, whilst extracting georeferenced points from VGI can only offer an impression of a user's intermediary places, appropriate computational processes can contribute towards a more contextual understanding.

2.3 Computational Approaches

2.3.1 Spatial Clustering

In their work, Lloyd and Cheshire (2017) used 2011 census data and tweets from December 2012 to January 2014 to investigate whether retail centre catchment areas can be detected from Twitter data. Whilst their datasets were temporally asynchronous, their use of adaptive kernel density estimation (KDE) to highlight unusual areas of activity within an otherwise unfiltered dataset was relatively successful in discovering unexpectedly large catchment areas. However, due to the limitations of their analysis they could not distinguish the direction of travel, thus potentially skewing the catchment areas with unrelated long distance travel. This could have been overcome by either incorporating temporal analyses to calculate if a user was travelling in or out, or by using NLP on the content of the tweet to extract relevant directional information. Nevertheless, Lloyd and Cheshire (2017) showed the ability for tweets to represent the narrative of habitual consumer-based mobility.

The limitations in over-analysing text are also shown in earlier works. Birkin et al. (2013) investigated behaviour patterns between deprived and more prosperous areas of Leeds, UK, using KDE and NLP to attribute single word summaries for wards (city council divisions of land). Whilst this was beneficial for the study in terms of aggregating semantics over large areas, it stripped out local context and thus barred them from discovering a more representative spatio-temporal narrative. Similarly, work by Hasan et al. (2013) used Foursquare and Twitter data to establish a hierarchy of popular activities and their associated locations using KDE and frequency metrics. However, their KDE method used a fixed bandwidth, undersmoothing unpopular areas and oversmoothing popular ones to create a more aggregated result. Whilst their categorical model had more motifs than Birkin et al. (2013), subsequently affording a slightly richer narrative representation, the aggregation and subsequent inability to analyse smaller pockets of activity contribute to under-represented local narratives within these wards. However, this is a general limitation of KDE, one that only a few academics have attempted to mitigate against.

Work by Steiger et al. (2015) similarly attempted a KDE aggregation model to correlate Twitter geolocation data with the UK census. Their one-word categories of "Home" and "Work" tweets, comprised of component topics, generally matched residential and commercial areas but at a very high level. However, their advanced statistical analyses produced a much more reliable model of mobility patterns than those previously mentioned. They applied the unsupervised topic detection algorithm latent dirichlet allocation (LDA) (Blei et al., 2003), KDE, Local Moran's I (Anselin, 1995) and Getis Ord Gi to build upon existing KDE methodologies. The latter two analyses assign values to the relationships between clusters of spatial observations and their centroids, as well as to local pockets of activity within a grid square. This allowed for a more statistical comparison between spatial and linguistic attributes of areas and highlighted whether or not nearby areas share similar topics and to what degree. Despite the intensive analysis, the aggregated topics were predominantly one-word, comprised of other one-word topics, thus similarly discarding narrative context.

3 Proposed Analytical Framework

As the importance of narrative inclusion has been argued, it is therefore necessary to construct a narrative-centric analytical framework. This will focus on extracting the rich, social information in combination with statistical methods rather than in competition. At three distinct stages motifs are extracted in parallel with statistical analyses that would otherwise have removed key contextual information. These two threads then combine to construct the narrative. An outline of the framework is presented in Figure 1.

3.1 Clustering Locations

To obtain sufficiently accurate spatial clusters, a combination of KDE models by Steiger et al. (2015) and Lloyd and Cheshire (2017) will highlight areas of activity including both local hotspots and the statistical differences between them considering the prevalent topics of discussion. This allows for detailed spatial and semantic comparisons, ideal for inferring general narrative themes within local areas.

3.2 Extracting Narratives

Reflecting on Tomashevsky (1965)'s definitions, extracting features from Twitter data is achievable through existing methods of tokenisation. These features can be constructed into motifs using topic modelling algorithms such as LDA. It is at this point that existing methods struggle; using computational methods to connote themes from motifs is challenging due to tokenisation removing vital contextual data. Therefore, after extracting the prevalent topics using LDA, the dataset can be searched for tweets matching the topics. Then, using part-of-speech tagging module to code each word as either a noun, adjective, adverb et cetera, noun phrases can be extracted that offer more description such as "bustling market". Furthermore, sentiment analysis can be simultaneously carried out on the messages, allowing for the understanding of how the topics are being discussed. The combination of these methods results in semantically rich topic descriptions from which themes can be extracted and narratives constructed.

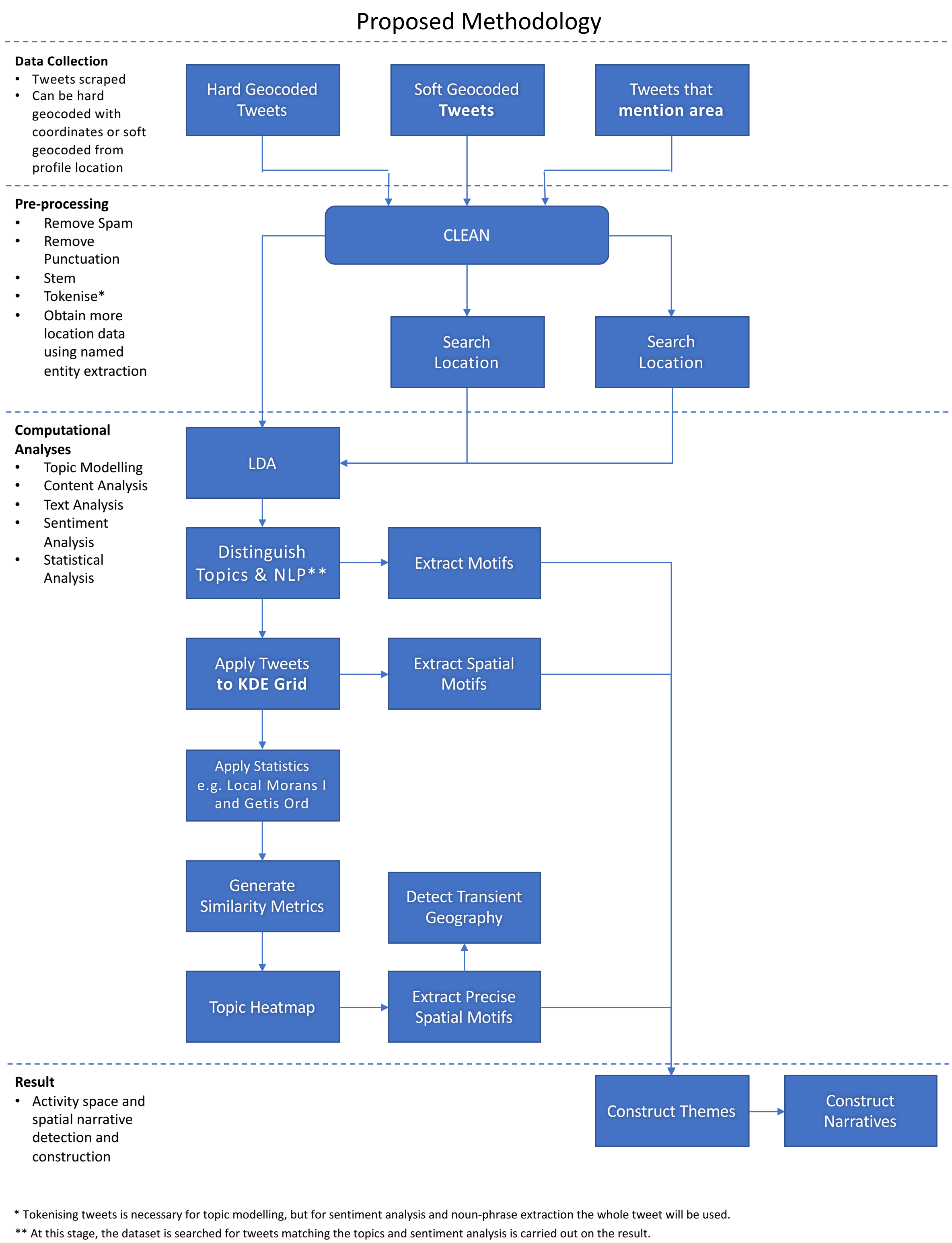

Figure 1: Flowchart depicting the proposed analytical framework.

4 Future Work

As this paper forms part of an ongoing PhD study into narrative extraction and activity space detection, future work will involve implementing this framework on a dataset of tweets gathered using the Southampton Web Observatory. This dataset, covering the whole of Hampshire, UK, will comprise of socially, spatially and temporally relevant Twitter data that will challenge this new framework.

References

Agnew, J. (2011). Space and place. In Agnew, J. and Livingstone, D., editors, *Handbook of geographical knowledge*, volume 2011, chapter 23, pages 316–330. Sage, London.

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2):93–115.

Birkin, M., Harland, K., and Malleson, N. (2013). The Classification of Space-Time Behaviour Patterns in a British City from Crowd-Sourced Data. *LNCS*, 7974:179–192.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(1):993–1022.

Bontcheva, K. and Rout, D. (2014). Making sense of social media streams through semantics: A survey. *Semantic Web*, 5(5):373–403.

Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research.

Farrow, E., Dickinson, T., and Aylett, M. P. (2015). Generating narratives from personal digital data: Using sentiment, themes, and named entities to construct stories. In *Human-Computer InteractionINTERACT 2015*, volume 9299, pages 473–477. Springer International Publishing.

Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. (2014). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*.

Gu, Y., Qian, Z. S., and Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67:321–342.

Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, page 1.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hulsey, N. and Reeves, J. (2014). The gift that keeps on giving: Google, ingress, and the gift of surveillance. *Surveillance and Society*, 12(3):389–400.

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

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. (2011). I'm the Mayor of my house: examining why people use foursquarea social-driven location sharing application. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, volume 54, page 2409.

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.

Maynard, D. and Hare, J. (2015). Entity-based Opinion Mining from Text and Multimedia. In Gaber, M. M., Cocea, M., Wiratunga, N., and Goker, A., editors, *Advances in Social Media Analysis*, chapter 4, pages 65–86. Springer International Publishing.

Mennis, J., Mason, M. J., and Cao, Y. (2012). Qualitative GIS and the visualization of narrative activity space data. *International Journal of Geographical Information Science*, 8816(March):1–25.

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.

Tamburrini, N., Cinnirella, M., Jansen, V. A. A., and Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84–89.

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