Towards a framework for automatic geographic feature extraction from Twitter

Enrico Steiger, Johannes Lauer, Timothy Ellersiek, Alexander Zipf

GIScience Research Group, Institute of Geography, University of Heidelberg, Berliner Straße 48 D-69120 Heidelberg
Email: {enrico.steiger; johannes.lauer, timothy.ellersiek,zipf}@geog.uni-heidelberg.de

1. Introduction

Interactive social media platforms offer a tremendous amount of volunteered, user-generated content (Flickr, Twitter, etc.). Together with volunteered geographic information (VGI) they potentially provide a valuable source of information which is increasingly recognized, but particularly in GIScience not utilized to its full potential. Twitter as one location-based social network in particular, provides the ability to sense geo-processes and to gain knowledge about the individual user perception towards geographic objects. A georeferenced tweet represents a proxy of a real world observation and contains spatial, temporal and semantic information. These social sensor measurements depend on particular tweet locations and are influenced by the individual user perception of urban space. Although there is a growing research body conducting Twitter analysis, a key challenge remains whether this noisy biased data source forms a representative sample for the knowledge discovery of geographic information. Location information retrieved from Twitter data is spatio-temporally and semantically uncertain. One of the main research aims is therefore to investigate whether geographic features from tweets can be detected and extracted. Furthermore, we explore whether the inferred geometries of features match with real world spatial objects (e.g. points of interest).

In this work we propose a framework to infer geographic features from unstructured georeferenced Twitter data using semantic topic modelling and spatial clustering techniques. Given the detected and extracted geographic features from Twitter, we applied a geometry computation and compared the results with map features from OpenStreetMap.

1.1 Related Work

There are a number of previous studies on a macroscopic scale aiming to infer direct or indirect geographic information from Twitter using provided metadata, the semantic tweet content or geographic coordinates. Cha et al. (2010) focus on enriching georeferenced tweets by inferring the location from user profiles and in addition their social network. Gonzalez and Chen (2012), Hiruta et al. (2012) and Lee and Hwang (2012) further develop a location inference system using user profile location, semantic classified tweet content or GPS coordinates from the geotag. Hong et al. (2012) develop a location aware topic model to correlate relationships between location and words. Dalvi et al. (2012) geolocate users by matching posted tweets containing indirect spatial information to real world spatial objects. Sengstock and Gertz (2012) introduce a framework for unsupervised extraction of latent geographic features from georeferenced Flickr data.

2. Methods

Tweets represent a spatio-temporal signal with a semantic information layer. We have extracted a semantic dimension over geographic space in order to infer geographical features on a small map scale (street level).

2.1 Dataset

For our case study we use a dataset only containing geotagged tweets from the area of Greater London. Table 1 shows some further details regarding the retrieved Twitter data.

Dataset	Greater London (UK)
Bounding Box (WGS 84)	-0.5543,51.2386,0.3038,51.731
Timespan	01/10/2013-31/03/2014
Covered Area	3265387 km²
Number of geotagged tweets	15.8 million
Number of tweeted User	433555

Table 1: Meta information for our selected Twitter dataset

2.2 Framework

All tweets are collected in real-time through the official Twitter streaming API (https://dev.twitter.com/docs/api/streaming). The semantic tweet content from every user is then preprocessed to remove whitespaces, punctuations and numbers. In the next step all tweet corpora from Twitter undergo a natural language processing step by applying tokenization, stemming and stop word filtering (Lewis et al. 2004). We are using latent dirichlet allocation (LDA) as one semantic probability based topic extraction model introduced by Blei et al. (2003). The unsupervised machine learning model identifies latent topics and corresponding word clusters from our large collection of tweets. This technique reduces the semantic dimensions and works efficiently especially on large unseen datasets. It is a sophisticated method compared to arbitrary simple keyword filtering techniques which have limited scalability. Figure 2 shows an exemplary LDA probabilistic topic extraction visualization for the highest assignments (>0.3) for the topic associated words “trafalgar” and “square”. The words “photo”, “london” and “england” also appear and show lower topic assignments (<0.3). As a result, high density areas of topic relevant classified tweets are closer to the real world object Trafalgar Square.

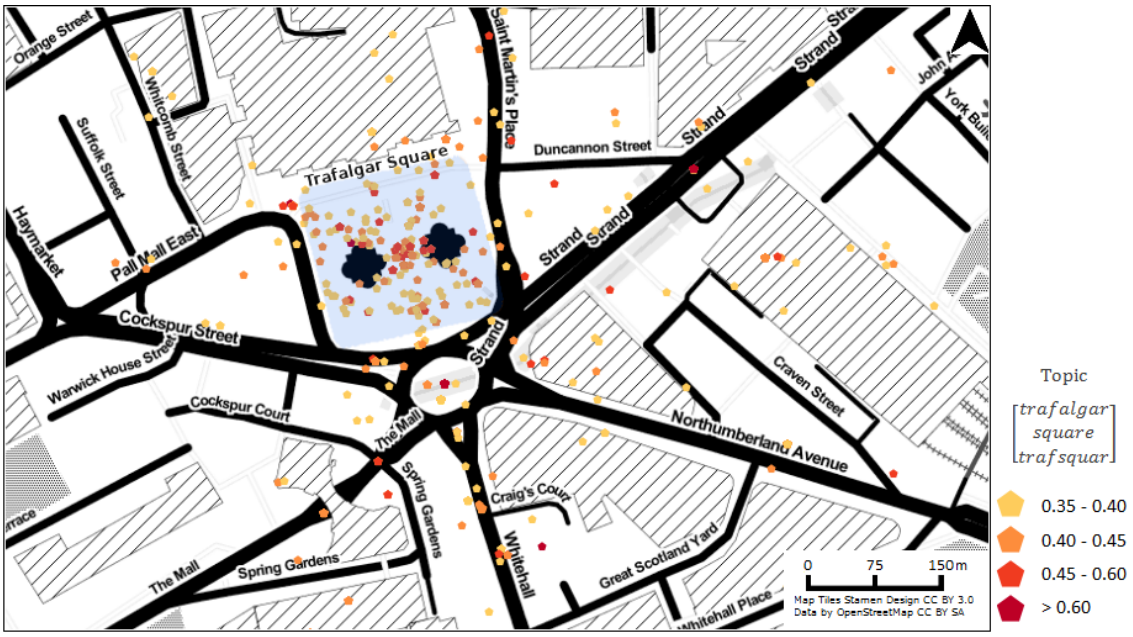

Figure 1: LDA topic association indicator for words “trafalgar” and “square” over all topic related filtered georeferenced tweets in London (n = 3796).

After the tweets have been processed and classified with LDA topic modelling, we chose DBSCAN (Ester et al. 1996) as a density based point clustering and classification algorithm to process the point cloud data. The algorithm detects dense clusters and filters noisy points. From the densest cluster where most tweets have been assigned to, we generate a trajectory which can be compared and matched with the corresponding geographic object from OpenStreetMap (OSM).

3. Results

3.1 Point Clustering

DBSCAN is applied in order to detect statistically significant semantic and geographic centroids of LDA classified tweets for the topic "oxford street". The Euclidean distances between the topic associated tweets have been normalized. The minimum number of points to form a cluster was defined to be 7 with a density reachability distance of ε=0.1. As a result (Figure 2) all tweets in cluster 1 (91% of total extracted features) are density connected points without scattering. This cluster is spatially concentrated along the real world geographical object Oxford Street. Associated tweets (n=1085) are our targeted point cluster. Cluster 2 and 3 are locally occurring dispersed clusters, showing a low density-reachable tweet distribution with a low amount of associated features.

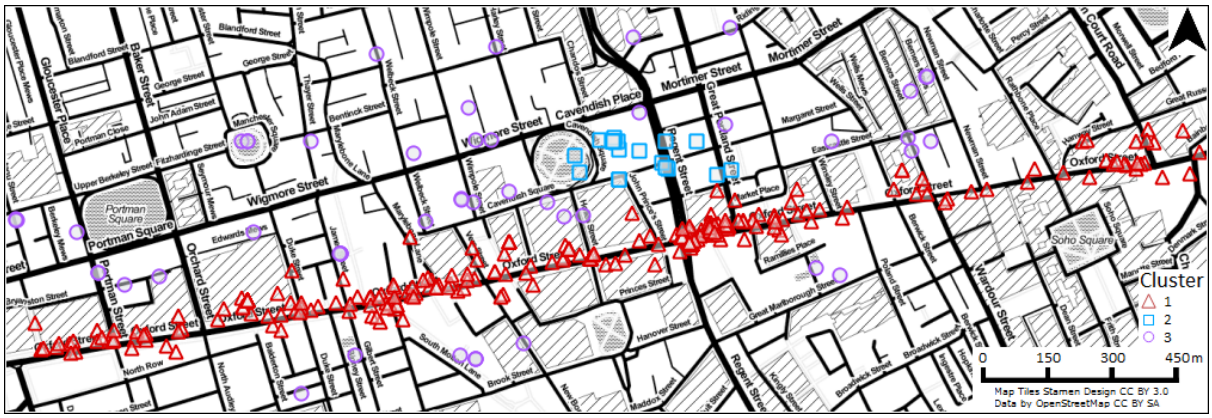

Figure 2: LDA topic associated words "oxford" and "street" for georeferenced tweets in London after applying DBSCAN clustering (n = 1186).

3.2 Geometry Extraction and OSM Feature Comparison

In order to extract geometric features and compare them with an existing map, several processing steps have to be taken. The first step is the extraction of the corresponding geometry for the new feature. In our case we created a linestring by applying the principal curve algorithm of Hastie and Stuetzle (1989), which is able to fit a line string to an unsorted point data set. The result is a geometric representation of Oxford Street (Figure 3). The second step is to match the new generated linestring with the corresponding feature from the OSM road network. As a quality indicator for the positional accuracy, the Hausdorff distance is calculated. For both linestrings, the Hausdorff distance is 0.0030468 which provides an indication for their similarity.

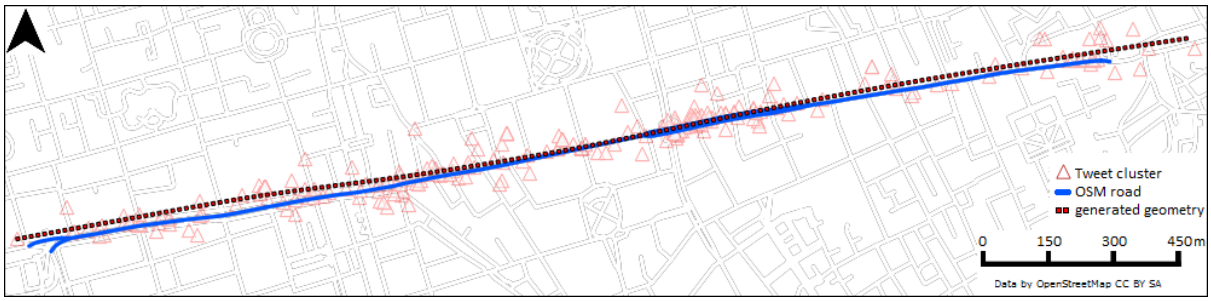

Figure 3: Oxford Street - extracted linestring geometry from tweets and comparison with OSM road (Hausdorff distance = 0.0030468)

4. Conclusion

Our results for the selected case study in London show that geographic features can be successfully extracted from Twitter by using geographic and semantic information. We were able to generate a new road feature from Twitter measurements which is quite similar to the mapped OpenStreetMap feature. Limitations of the study include the geographic objects themselves which might be too complex to be clearly detected from the spatial-semantic signal, or the tweet signal might not be significant enough and too sparse to be detected at all.

References

Blei, D., Ng, A. and Jordan, M., 2003, Latent dirichlet allocation. *Journal of machine Learning research*, 993–1022.

Cha, Meeyoung, et al., 2010, Measuring User Influence in Twitter: The Million Follower Fallacy. In: *ICWSM 10*, 10-17.

Dalvi, N., Kumar, R. and Pang, B., 2012, Object matching in tweets with spatial models. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*. New York, USA, 43.

Ester, M. et al., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96.

Gonzalez, R. and Chen, Y., 2012, TweoLocator: A Non-Intrusive Geographical Locator System for Twitter. In: P*roceedings of the 5th International Workshop on Location-Based Social Networks*, 24–31.

Hastie, T. and Stuetzle, W., 1989. Principal Curves. Journal of the American Statistical Association, 84(406), 502–516.

Hiruta, S. et al., 2012, Detection , Classification and Visualization of Place-triggerd Geotagged Tweets. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*.

Hong, L. et al., 2012, Discovering geographical topics in the twitter stream. In: *Proceedings of the 21st international conference on World Wide Web - WWW '12*. New York, 769.

Kinsella, S., Murdock, V. and Hare, N.O., 2011," I ' m Eating a Sandwich in Glasgow ": Modeling Locations with Tweets. In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 61–68.

Lee, B. and Hwang, B.-Y., 2012, A Study of the Correlation between the Spatial Attributes on Twitter. *2012 IEEE 28th International Conference on Data Engineering Workshops*, 337–340.

Lewis, D.D. et al., 2004, RCV1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research*, 5, 361–397.

Sengstock, C. and Gertz, M., 2012, Latent geographic feature extraction from social media. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*. New York, USA, 149.