Uncovering Latent Mobility Patterns from Twitter During Mass Events

Enrico Steiger, Timothy Ellersiek, Bernd Resch and Alexander Zipf
GIScience Research Group, Heidelberg University/Germany · enrico.steiger@geog.uni-heidelberg.de

Full paper double blind review

Abstract

The investigation of human activity in location-based social networks such as Twitter is one promising example of exploring spatial structures in order to infer underlying mobility patterns. Previous work regarding Twitter analysis is mainly focused on the spatiotemporal classification of events. However, since the information about the occurrence of a general event can in many cases be considered as given, one identified research gap is the exploration of human spatial behavior within specific mass events to potentially characterize underlying, locally occurring mobility clusters. One key challenge is to explore whether this noisy biased dataset can be a reliable source for the knowledge discovery of human mobility during mass events. In this paper we therefore present an advanced methodologycal framework, including a generative semantic topic modeling and local spatial autocorrelation approach, to observe both spatiotemporal and semantic clusters during a major sports event in Boston in the US. Our results of the observed spatiotemporally and semantically clustered tweets within the selected case study area have shown the possibility of deriving intra-urban event related mobility patterns with similar spatiotemporal movement.

1 Introduction

The growing number of handheld devices equipped with a range of sensors has created new possibilities to infer latent mobility patterns from crowdsourced information. With ubiquitous access to broadband Internet and means to utilize positioning systems such as GPS, spatiotemporal knowledge from digital footprints can be extracted. With the influx of Web 2.0 technologies and services, individuals are able to participate in collaborative networks such as Wikipedia or OpenStreetMap without requiring prior expert knowledge. The latter of course is probably one of the most prominent applications of crowdsourced geodata, which, amongst others, also led to the notion of Volunteered Geographic Information (VGI) (GOODCHILD 2007).

Online social networks are also a major part of this development. These microcosms reflect many people's lives in data, and they pose immense value, both economically and scientifically. Location Based Social Networks (LBSN) further enhance existing online social networks, adding a spatial dimension by utilizing location-embedded services. Whether users are uploading geotagged photos via Flickr or Instagram, checking in at a venue with Foursquare, or commenting on a local event via Twitter − personal locations become a key

GI_Forum – Journal for Geographic Information Science, 1-2015.

point of interaction (ZHENG 2011). In this work we focus on Twitter data containing spatial attributes to find latent mobility patterns.

One outcome of a previously conducted systematic literature review revealed that researchers are predominantly investigating event detection from Twitter. However, since in many cases the information about the occurrence of a general event can be considered as given (e.g. for some mass events like concerts etc.), it seems that there is currently an overly strong concentration of studies (STEIGER et al. 2015). Hence, the research should rather be focused on the exploration of human spatial behavior (MILLER & GOODCHILD 2014) on a larger, intra-urban scale level, in order to gain knowledge about the underlying geographic processes within specific events. One potential application of this research can be seen in the ability to detect similar spatial human mobility patterns within events in a (near) real-time manner, and thus may add to or validate existing information sources.

Therefore, as one main research question (RQ1) of this paper, we analyse georeferenced tweets for a selected case study to investigate whether similarities among spatiotemporal and semantic information reveal intra-urban human mobility activities during mass events. Furthermore, we examine whether the observed clusters are a proxy indicator for the characterization of underlying urban structures (RQ2). As one part of the results we present a novel visual analytics approach for characterizing mobility patterns from Twitter during mass events.

2 Background

The dataset used in this analysis is retrieved from Twitter, where users can post so called tweets comprising short text messages with up to 140 characters. With the granted authorization of each user set in the mobile device, tweets may also contain a spatial attribute featuring the specific GPS position at the time of the actual post. Therefore, we can obtain three information layers from tweets, in particular the semantic data layer embodying the content of the message and the spatiotemporal signal including the location and time of the specific message. One main task is to find repeating patterns by combining all dimensions since, e.g. the semantic layer doesn't necessarily have to align with the spatial component of the same tweet; unlike in Foursquare, where users can "check in" at specific venues (restaurants, hotels etc.), which already characterize categories of activities, we do not have any a priori knowledge regarding certain human social activities in Twitter.

Thus, information retrieved from Twitter data is spatiotemporally and semantically uncertain. Most importantly, one must be aware that the accuracy of locational information from tweets can be influenced by mobile device characteristics, urban environments or other factors such as the GPS dilution of precision. The semantic layer detected within an event in time may correspond to past or future events. Additionally, the text corpora derived from tweets are relatively vague, including ambiguous semantic information, which might be only a weak indicator of a real world event. As users do not post tweets equally distributed in geographic space and time, one must also consider the heterogeneous dispersion of tweets. Furthermore, georeferenced tweets only represent a small fraction of the overall tweets available.

2.1 Related Work

In this subsection we will briefly introduce a selection of related research studies. Thus, the number of mentioned studies does not claim to be exhaustive. The investigation of spatial, temporal and semantic tweet frequencies and patterns from Twitter as a social sensor in order to detect real world events, has been conducted in a number of studies (SAKAKI et al 2010, LEE & SUMIYA 2010, CHAE et al. 2012).

Several researchers solely focus on the semantic attribute of tweets to detect events from Twitter by using natural language processing. BECKER & GRAVANO (2011) and JACKOWAY et al. (2011) for example identify real-world events and news content on Twitter by extracting and classifying topics using term frequency analysis and naive bayes classifiers. Using spatiotemporal and textual information from Twitter, researchers aim to discover events in the area of disaster- (SAKAKI et al. 2010), crisis- (STEFANIDIS et al. 2011), and mobility management (RIBEIRO et al. 2012). In the latter application, a number of studies aim to infer general human mobility using Twitter. LEE & SUMIYA (2010) study user behavior patterns in Twitter by measuring geographic regularities and detecting geo-social events within regions of interest (ROI). BOETTCHER & LEE (2012) classify events based on different geographical scales by counting average daily keyword frequencies over space, utilizing the DBSCAN clustering algorithm, and, furthermore, group terms according to their relevance to a local event. CHAE et al. (2012) examine social ties and their impact on human mobility patterns using mobile network check-in information and cell phone areal information.

3 Methodology

We argue that coincident social activities may uncover underlying latent human mobility patterns, as these footprints include observations at a given time for a certain place. For instance, a number of Twitter posts recorded at the same event could constitute a cluster, since they are highly spatially, semantically, and temporally close to each other. The proposed analysis framework is able to retrieve, store, and process Twitter data in a (near) real-time manner in order to detect these clusters. First of all, tweets are collected in real-time through the official Twitter streaming API[1]. The conducted workflow shown in Figure 1 comprises three main steps after data retrieval: data pre-processing, semantic similarity measurement, and local spatial autocorrelation analysis. Afterwards the derived clusters are visually explored within a WebGL framework.

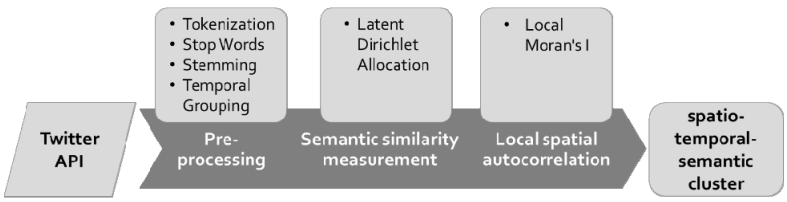

Fig. 1: Analysis framework

[1] https://dev.twitter.com/docs/api/streaming

3.1 Preprocessing

All Twitter posts undergo a pre-processing step to reduce the semantic dimension of textual information. Initially, whitespaces and punctuations are removed from every tweet before the remaining word vectors are natural language processed by applying tokenization, stemming, and stop word filtering. After raw textual information from Twitter posts are divided into single words through tokenization, most commonly occurring stop words are filtered out. These resulting words have the same probability of appearing in another randomly selected document, and therefore reduce the amount of noise. The advantage and performance of this approach has been described by METKE-JIMENEZ et al. (2011). The remaining word vectors are the input for the following semantic similarity measurement. After pre-processing the textual features from tweets, we classify the temporal component (time of each post) in order to assess temporal similarity from tweet activities. Time bins covering every hour (24 bins) and for each day are therefore created as a categorical variable in order to group tweets when sharing similar activities in a geographical proximity and close in time. In the last pre-processing step, all selected tweets are normalized over each user to avoid an over- or underrepresentation of particular Twitter users due to varying tweet frequencies.

3.2 Semantic Similarity Measurement

With the final goal of assessing semantic similarity for the collected tweets, we are utilizing latent dirichlet allocation (LDA) as one semantic probability based topic extraction model introduced by BLEI et al. (2003). The unsupervised machine learning model detects latent topics and corresponding word clusters from our large collection of tweets and has been successfully applied in previous studies (PAN & MITRA 2011, FERRARI et al. 2011, KLING et al. 2012). The hierarchical Bayesian model clusters co-occurring words into topics (bag-of-words model) and performs efficiently, particularly on large unseen datasets by using a given training set. Compared to simple keyword filtering techniques with a limited scalability (JACKOWAY et al. 2011), the LDA model is also capable of distinguishing and assigning similar phrases with different context into separate topics. LDA assumes that documents (in our case tweets) contain a random number over latent topics per document, where each topic is characterized by a distribution over words. One main challenge when utilizing LDA is the posterior parameter estimation and computation of variables such as the number of topics k. For the LDA parameter inference, we therefore use Gibbs sampling, which is based on a Markov chain Monte Carlo method. This sampling from probability distributions solves a key inferential problem and optimizes the topic model parameter selection process following GRIFFITHS & STEYVERS (2004).

3.3 Local Spatial Autocorrelation

To assess whether observed nearby tweets cover the same topics and show similar associated topic indicator values (or not), we apply local measurements of spatial association. Spatial association is indicated by the presence of spatial autocorrelation. With Local indicators of spatial association (LISA), introduced by ANSELIN (1995), we measure the extent to which statistically inferred spatial associations depart from the null hypothesis, being complete spatial random observations. The deviating observations are identified locations of spatial clusters and spatial outliers denoting our targeted human social activities. *Local*

Moran's I_i test statistic determines the deviations of a value in an observation area i which belong to the specified neighbourhood set J_i, where W_{ij} is the spatial weight of i's and j's location and $\bar{x}$ is the mean of variable x (in our case probabilistic LDA topic association indicator).

$$I_i = (x_i - \bar{x}) \Sigma_{j \in J_i}^{n} W_{ij} (x_j - \bar{x})^2 \quad (1)$$

$$Z(I_i) = \frac{I_i - E(I_i)}{\sqrt{V(I_i)}} \quad (2)$$

A positive value for I shows positive spatial autocorrelation with features having similar neighbouring features of high or low attribute values. Analysing the resulting $Z(I_i)$-scores and applying different significance tests we can distinguish between local spatial autocorrelation clusters of high values (HH) indicating a hotspot, low values (LL) indicating a cold spot, high values surrounded by low values (HL), and low values surrounded by high values (LH) indicating a spatial outlier.

4 Results

The previously described methods from the analysis framework (3) have been applied to our selected case study. The results are presented in the following section.

4.1 Case Study

For our case study we use a dataset only containing georeferenced tweets from the area of Greater Boston over one year (Table 1). In this timeframe we looked for topics associated with the baseball world series, and generated a temporal subset with the highest peaks. The subset used for our analysis contains 251,771 posted tweets during a major baseball event in Boston (including 5 additional days before and after the corresponding event).

Table 1: Meta information of used social media data for selected case study

Dataset	Greater Boston (USA)
Bounding Box (WGS 84)	-71.284, 42.191,-0.8113, 42.5509
Timespan	31/07/2013-31/07/2014
Number of geotagged tweets after pre-processing	3,2 million
Number of individual users	186,395

4.2 Results Semantic Similarity Assessment

When focusing on the temporal-semantic distribution of LDA probabilistic extracted topics for the highest assigned topic associated words over all observed tweets (Figure 2), anomalous frequency peaks during the sports events can be detected.

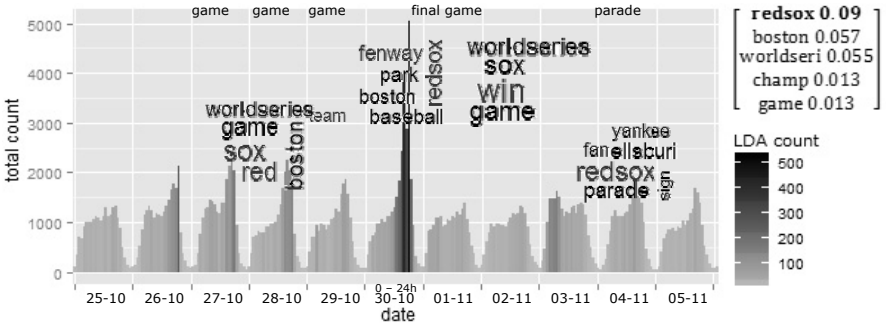

Fig. 2: Temporal-semantic frequencies of all tweets and for the LDA classified topic (grey to black highlighted) during the sports event. The most frequently occurring words denoted by the label size during the given time intervals are shown above the corresponding bars.

The periodically repeating daily signal of observed tweets shows less autocorrelation, and differs during these series of games and the following street parade. Note that the amount of topic associated tweets and the most frequently occurring terms, increases during the specific games and reaches its climax on the 30th November 2013 with the final game. The highest topic assigned word "redsox" within the LDA model (which is Boston's baseball team name) is selected, as it is the most dominant word for describing our given set of topics. Because the LDA model embodies textual documents as a mixture of topics, these probability based assigned words are the most likely to have generated the original text corpora. For visualization purposes we only show the 5 highest topic assigned words. The majority of tweets during the final game were posted from 8pm onwards until midnight. The hourly frequency of LDA semantic classified thus temporally corresponds to the real world event.

4.3 Results Local Spatial Autocorrelation

Since we previously focused on extracting the temporal-semantic information from tweets to identify clusters of social activities, we are now considering spatial aspects of the data. With Local Moran's I we can measure the degree of spatial association for all topic classified tweets in relation to a single tweet observation. The generated maps in Figure 3 visualize a time series of pre-event, event, and post-event observed local spatial clusters with a significant positive autocorrelation indicating event related activities. Clusters of high values coloured red (HH) and clusters of low values coloured blue (LL) having a Local Moran statistic p-value below 0.01, are shown on the maps. Very high or low Local Moran's I z-scores within the given confidence level (99%) above 2.58 z-score are unlikely to be the result of random chance, and are statistically significant clusters of similar or dissimilar values (in our case the LDA association indicator of covering event related textual information). Therefore, we can reject the null hypothesis for our observations being complete spatial randomness (CSR). For visualization purposes we only show a time series including the pre- and post-phase of the event to demonstrate the ability of detecting evolving latent mobility related clusters. Having a closer look at the spatiotemporal distribution of event associated tweets before the game, high topic intensities with a positive autocorrelation

spatially cluster near major public transport hubs (i.e. North Station/World Trade Center Station), squares (i.e. Faneuil Square/ Harvard Square), and within the nearby surroundings where the sports event takes place. The highest positive Local Moran's I z-scores during the game between 8-10 pm, indicating high event (topic) associated tweets, concentrate inside the event venue (baseball stadium). The previously detected cluster areas around public transport hubs (i.e. North Station/World Trade Center) before the event, now show less semantic topic affinity (LL), and are low value clusters. Tweets posted shortly after the game (10-12 pm) cluster with a high positive autocorrelation within and around the sports stadium, but are also more dispersed within the whole city of Boston and Cambridge. The wider distribution of positively spatial autocorrelating clusters densely concentrates within the proximity of main public squares.

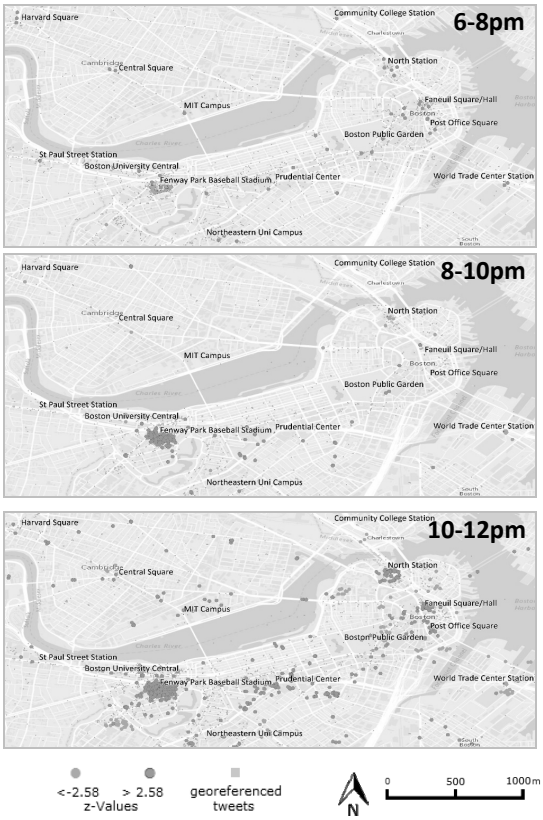

Fig. 3: Time series of derived local spatial clusters of event related topics with significance test applying Local Moran's I (base map: Canvas/Esri, HERE, DeLorme, MapmyIndia, Data by OpenStreetMap CC BY SA)

4.4 Explorative Visual Analysis of Results

In order to facilitate cluster detection, we integrated a visual exploration tool to analyse spatiotemporal and semantic patterns of events within an interactive space-time cube fol-

lowing GATALSKY et al. (2004). Using the Three JS framework, the clustered Twitter point features are plot into a three-dimensional space. The constructed WebGL[2] scene in Figure 4 consists of an underlying aerial image from Fenway Park stadium, featuring the geographic position of each tweet with the categorized hourly time bins as a further spatial dimension (z-axis). In addition, the results of the semantic similarity measurement (3.2) and the local spatial autocorrelation (3.3) are added as attribute values for each point feature, which can be individually selected within the layer. When comparing all dimensions, including the feature's attributes, one can discern the spatiotemporal and semantic extent of the clustered tweets, uncovering an inflow behaviour. Tweets posted during the sports event (temporal proximity) within the stadium (geographical proximity), and containing similar textual information referring to the baseball game (semantic proximity), highly autocorrelate with each other.

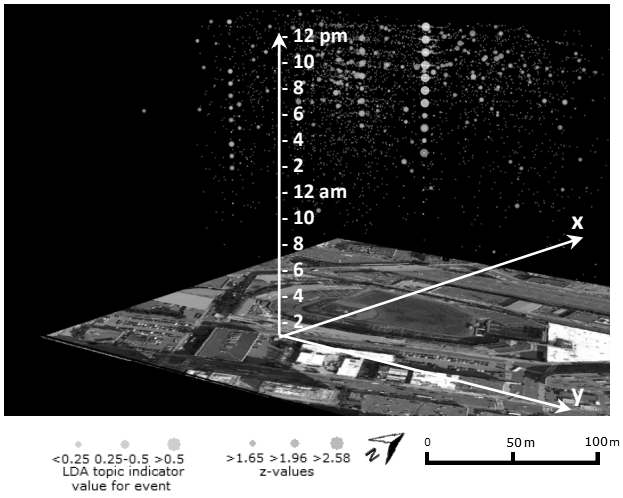

Fig. 4: Space-time cube showing LDA semantic classified tweets (orange) related to the sports event over all tweets (white dots) with a positive Local Moran's I z-value denoted in light violet (base map: Bing Aerial Map 2010 Microsoft Corporation)

5 Conclusion and Future Works

In this paper we described a spatiotemporal and semantic analysis framework for georeferenced tweets, in order to investigate latent intra-urban human mobility patterns during mass events. The results of our selected case study in Boston show that by combining all temporal, geographic, and semantic information from tweets, and by applying semantic and spatial analysis methods, we are able to infer event related activities as an indicator of peoples underlying mobility behaviour (RQ1). To a certain extent the analysis revealed footprints indicating intra-urban mobility behaviour moving towards an event venue and vice versa. These event-related patterns have shown spatial clustering in the vicinity of central squares

[2] http://koenigstuhl.geog.uni-heidelberg.de/opentrafficflow2/

and major transportation hubs, and are therefore a proxy of characterizing urban infrastructure during mass events (RQ2). This research along with the methodological approach can therefore potentially be applied in other regions with limited access to official data and knowledge about urban mobility processes during mass events. The possible application of this research can also be seen in the ability to detect mobility patterns during mass events in a (near)-real time manner, helping to provide stakeholders and decision makers with a better understanding of mass movement processes. In one future piece of work we aim to expand our study across several regions, and for different types of events to compare detected mobility patterns with each other.

Finally, the conducted study has some limitations. The performance of semantic reduction techniques (3.1) and natural language processing varies due to the uncertainty of Twitter posts, considering domain specific writing styles and unpredictable Internet oriented terms. The efficiency of LDA topic modelling (3.2) depends on the initial parameter inference (e.g. number of topics) assuming to have a probabilistic dirichlet topic distribution. We applied spatial autocorrelation analysis (3.3) to detect clusters that show a semantic similarity over time and space. However, since Local Moran's I statistics measure the degree of spatial association for an observed value assuming a normal distribution, with respect to spatial stationarity and variance effects (ORD & GETIS 2001), these model assumptions do not explain the distribution of every geographic phenomena.

One clear limitation of the current methodology for extracting information is the reliance on Twitter data; assuming that Tweets are written in-situ, referring to an event at a location and time they have been published. The derived spatiotemporal and semantic signals from Twitter also might not be significant enough to serve as a proxy indicator of characterizing complex mobility behaviour.

Acknowledgments

This research was funded through the graduate scholarship program "Crowd-analyser − spatiotemporal analysis of user-generated content" supported by the state of Baden Wurttemberg. We also thank the existing social network community of Twitter for providing access to free available social media data.

References

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

BECKER, H. & GRAVANO, L. (2011), Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM 11, 438-441.

BLEI, D., NG, A. & JORDAN, M. (2003), Latent dirichlet allocation. Journal of Machine Learning Research, 993-1022.

CHAE, J., THOM, D., BOSCH, H., JANG, Y., MACIEJEWSKI, R., EBERT, D. S. & ERTL, T. (2012), Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), 143-152.

GATALSKY, P., ANDRIENKO, N. & ANDRIENKO, G. (2004), Interactive Analysis of Event Data Using Space-Time Cube. In: Eighth International Conference on. IEEE.

GETIS, A. & ORD, J. K. (1992), The Analysis of Spatial Association. Geographical Analysis, 24 (3), 189-206.

GOODCHILD, M. (2007), Citizens as sensors: the world of volunteered geography. GeoJournal, 69 (4), 211-221.

GRIFFITHS, T. & STEYVERS, M. (2004), Finding scientific topics. In: Proceedings of the National academy of Sciences of the United States of America, 101, 5228-5235.

HECKMAN, J. J. (1979), Sample Selection Bias as a Specification Error. The Econometric Society Stable, 47 (1), 153-161.

JACKOWAY, A., SAMET, H. & SANKARANARAYANAN, J. (2011), Identification of live news events using Twitter. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks – LBSN '11, 248-260.

KLING, F., KILDARE, C. & POZDNOUKHOV, A. (2012), When a City Tells a Story . Urban Topic Analysis, 482-485.

KOSALA, R. & ADI, E. (2012), Harvesting Real Time Traffic Information from Twitter. Procedia Engineering, 50 (Icasce), 1-11.

LEE, R. & SUMIYA, K. (2010), Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Location Based Social Networks – LBSN '10, 1.

METKE-JIMENEZ, A., RAYMOND, K. & MACCOLL, I. (2011), Information extraction from web services: a comparison of Tokenisation algorithms.

MILLER, H. J. & GOODCHILD, M. F. (2014), Data-driven geography. GeoJournal.

ORD, J. K. & GETIS, A. (2001), Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation. Journal of Regional Science, 41 (3), 411-432.

PAN, C.-C. & MITRA, P. (2011), Event detection with spatial latent Dirichlet allocation. Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries – JCDL '11, 349.

Ribeiro, S. S. Jr. et al. (2012), Traffic Observatory: a system to detect and locate traffic events and conditions using Twitter. In: Proceedings of the 5th Workshop on Location-Based Social Networks, 5-11. DOI:10.1145/2442796.2442800.

SAKAKI, T., OKAZAKI, M. & MATSUO, Y. (2010), Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, 851-860.

SENGSTOCK, C. & GERTZ, M. (2012), Latent geographic feature extraction from social media. Proceedings of the 20th International Conference on Advances in Geographic Information Systems – SIGSPATIAL '12, 149.

STEFANIDIS, A., CROOKS, A. & RADZIKOWSKI, J. (2011), Harvesting ambient geospatial information from social media feeds. GeoJournal, 78 (2), 319-338.

STEIGER, E., DE ALBUQUERQUE, J. P., ZIPF, A. (2015), An Advanced systematic literature review on spatiotemporal analyses of Twitter data. Transactions in GIS.

ZHENG, Y. (2011), Location-based social networks: Users. Computing with Spatial Trajectories. Computing with Spatial Trajectories, 243-276.