GI_Forum 2018, Issue 1
Page: 243 - 250
Short Paper
Corresponding Author:
barbara.pernici@polimi.it
DOI: 10.1553/giscience2018_01_s243

Talking about Places: Considering Context for the Geolocation of Images Extracted from Tweets

Chiara Francalanci[1], Barbara Pernici[1], Gabriele Scalia[1] and Gunter Zeug[2]

1 Politecnico di Milano, Italy
2 Terranea, Germany

Abstract

This paper investigates the extraction of geolocated images from social media. Pictures taken with a mobile device are typically georeferenced, but social media may or may not provide geo-coordinates, depending on their privacy policies. Our goal is to geolocate images extracted from Twitter to support emergency services in natural disasters. As the number of tweets with native georeferences is limited, we introduce algorithms that take advantage of various contextual clues included in social media posts to help increase the proportion of posts that can be geolocated. Using an explorative approach, we also investigate how to locate, in other social media, images that were originally embedded in tweets. The application of these context-based algorithms to a case study is discussed.

Keywords:

tweet geolocation, context modelling, emergency mapping

1 Introduction

The use of social media in emergency contexts has been studied extensively in the literature (e.g. Castillo, 2016; Imran et al., 2015). The most widely addressed research issues are the automated detection of new event locations (Sakaki et al., 2013), and the assessment of the impact of an event by means of hot and cold spot analysis (Resch et al., 2018). These research issues require geolocated information, which is recognized as a limiting factor, as a geotag is included in only 0.3% to 3% of all Twitter posts (de Albuquerque et al., 2015).

Several authors have investigated tweet geolocation (Jurgens et al., 2015); in general, the focus is on identifying the most frequent location of the author of the post, in terms of specific places, such as his/her home town. Recently, research has shifted focus to derive more precisely geolocated information (Middleton et al., 2014; Paraskevopoulos and Palpanas, 2016; Francalanci et al., 2017), such as specific points of interest (POI) (e.g. particular streets, named buildings or monuments), associated with precise geographical coordinates.

Consistent with this shift in focus, our main research goal is to extract from tweets images that are potentially useful for emergency mapping, and to associate them with a geolocation that is precise enough to support rapid mapping. We have developed algorithms that infer a geolocation from the information that characterizes a post both directly (i.e., the post content and its metadata), and indirectly (i.e., describing how the post is connected with other content through its author's interactions with other authors or links to other social media). A detailed and formal presentation of the algorithms can be found in Scalia (2017).

In Section 2, the principles of the context-based algorithms are presented. In Section 3, the methodology for assigning locations to extracted media is described; explorative techniques to further increase recall and accuracy are also discussed. In Sections 4 and 5, we evaluate our approach with a case study.

2 Context-based geolocation algorithms

Social media posts are typically difficult to disambiguate, especially if taken in isolation. A sentence in a traditional document is easier to interpret as the overall document provides a *context* for disambiguation. For example, a document that deals with agriculture is not likely to use the word *apple* to refer to a company. However, social media posts lack this context. Our focus is on posts that link images, as multimedia information is particularly useful for mapping purposes.

We make a distinction between *georeferenced posts*, i.e., posts that are natively associated with geocoordinates, and *geolocated posts*, i.e., posts that are associated with geolocation information by the algorithms proposed (Scalia, 2017). The levels of precision for geolocations are borrowed from the categorization of OpenStreetMap (Haklay and Weber, 2008). Levels equal to or above admin_level = 6 are sufficient for aerial images, since they refer to a broader area, while levels 10–15 (the maximum) are more suitable for identifying precise locations.
The idea underlying our algorithms is to search for *text references* about locations in the text of the posts (extracting Named Entities (NEs) and using a gazetteer), and to disambiguate and validate them by building a context that acts as a *reinforcement* for the candidate locations.

The context for a post is built in two steps: a *local* step (*CIME local*) which exploits the information that can be extracted from the post itself, and a *global* step (*CIME global*) which adds the information gathered from the social and conversational network of the post. The global step is employed only when the local step cannot disambiguate the candidate locations reliably. In particular, as illustrated in Figure 1:

The *local context* is derived from the post's text in combination with all the extracted candidate locations, hashtags and metadata.

- The *global context* adds the locations previously disambiguated in other posts that are connected to the target post: belonging to the same conversation, posted by authors who interacted recently, posting the same (or very similar) media, or sharing the same hashtags.
  The context is used to *rank* candidate locations, based on the admin_levels for

locations and their relationships. To do this, mutual hierarchical and spatial relationships are taken into account. The goal of this ranking is twofold:

- Disambiguating, i.e. choosing as the most likely location the one with the highest reinforcement level.
- Ensuring a certain confidence level for the disambiguated locations, by selecting the most likely location only if its rank score is higher than a particular confidence threshold.

Note that it is possible for several candidate locations related to the same text entity or different text entities to have the same score (a score that is higher than the threshold); in this case, multiple locations are assigned to the same post.

The global disambiguation step allows the *propagation* of disambiguated locations over the social network, since it can be performed recursively using previously disambiguated posts to disambiguate other ambiguous posts.

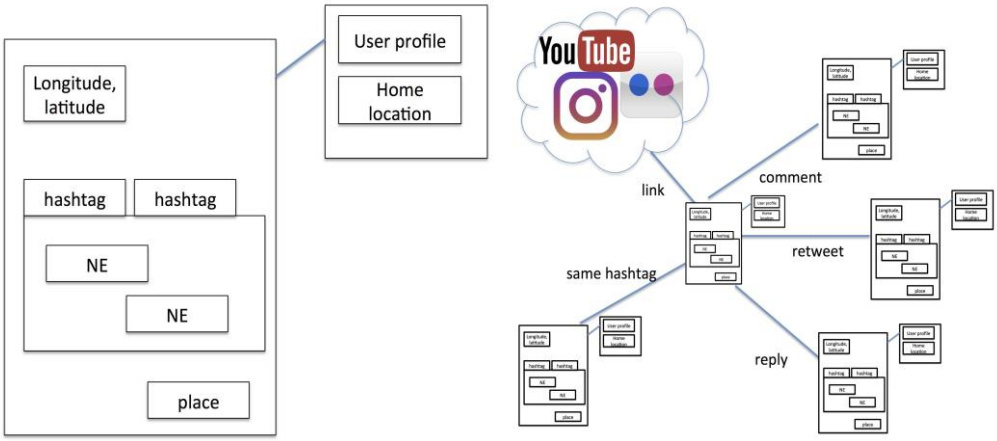

Figure 1: Elements of the local (left) and global (right) contexts

3 Media geolocation and exploration

The output of the geolocation algorithm presented in the previous section is a set of locations assigned to posts for which enough confidence and precision could be provided. The locations assigned to posts are transferred to their attached media. The rationale behind this choice is that, the geolocation algorithm being content-based, the output locations have a high probability of being related to the post's content (i.e., what people are talking about) and, therefore, to the post's media. Indeed, in some cases the geolocation provided by the algorithm is more accurate than the native georeferences of the posts, since the latter are related more to the *posting location* than to the *content location*.

The media geolocated by this approach are both those *directly* attached to tweets and those *indirectly* linked by tweets through linked media: Instagram, Flickr, YouTube and Facebook.

In this respect, media geolocation provides a basis for *refinement* and *exploration* in several ways:

- Same media attached to different posts. This requires *tracking down* the extracted media that have to be merged and integrating the locations extracted from the related posts.
- Finding new media related to the geolocated one. As posts are connected with each other, corresponding media are also connected, and the geolocation associated with a media could be exploited for other related media.
- Gathering and exploiting the information associated with the media themselves to further refine geolocation, in terms of precision or accuracy.

This third possibility is especially useful when considering media that are embedded in links to other social media, which may provide additional information, such as a title or description, to improve geolocation.

The approach was developed in Python, using Stanford CoreNLP (Manning et al., 2014) for Named Entity Resolution and OpenStreetMap with Nominatim (Haklay and Weber, 2008).

4 Analysis of a case study

The case study presented here is related to the Copernicus Emergency Mapping Service (EMS) activation for the 2014 Southern England floods[1]. For this activation, we extracted tweets for the period 10–15 February 2014 that related to two of the mapped areas (Bridgeport and Maidenhead); for these areas, the first EMS maps after the activation were produced on 12–13 February 2014. In Figure 2, we show the two areas of interest for the event, corresponding to two emergency maps created within EMS. The first aspect to be noted is the extremely low number of natively referenced tweets, which are indicated in Figure 2: only four for the two areas combined, obtained by crawling Twitter for the keywords 'floods' and 'england floods'. The number of georeferenced tweets for the 2014 Southern England flood case study is 3%, in line with results mentioned in the literature. Only 310 georeferenced tweets (out of 108,575) contain images (0.2%) for the areas in the crawling period.

We focus on supporting rapid mapping activity with images extracted from tweets, geolocated through the algorithms presented above.

In Table 1, for each geolocation strategy – a) using the native georeference of the tag if available, b) local context, and c) global context – we show the number of tweets containing images in the two areas. The table shows that the number of tweets that can be exploited for the purposes of rapid mapping can be increased considerably, from 4 to 45, through the context-based geolocation algorithm. The Table also shows that the number of tweets that can be useful for rapid mapping is statistically significant (more than 60% in all cases). The relevance of the media for rapid mapping activities was checked by annotating the geolocated tweets manually.

[1] http://emergency.copernicus.eu/mapping/list-of-components/EMSR069

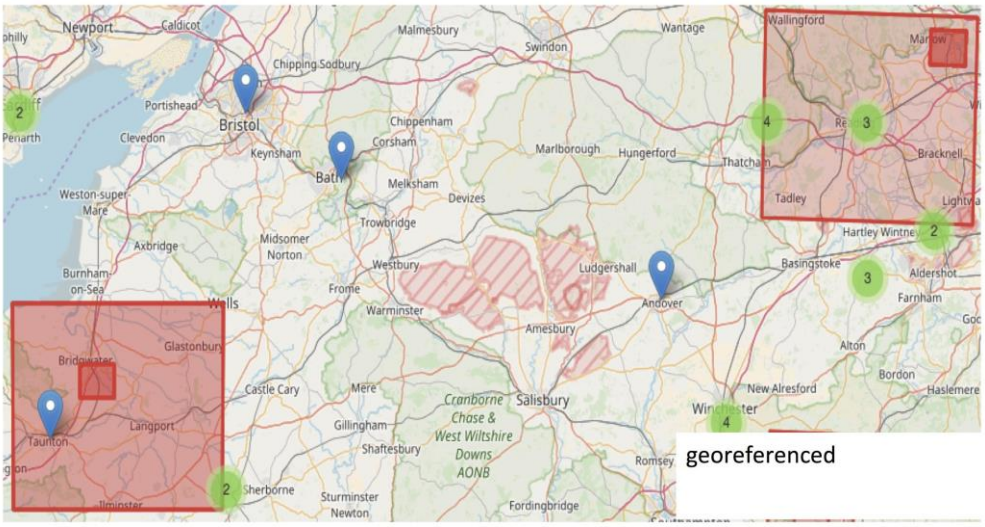

Figure 2: Georeferenced tweets in two mapped areas (in red) in the 2014 Southern England floods

In general, from the analysis of our case study concerning relevance, we confirmed the finding of Albuquerque et al. (2015): tweets in areas closer to the target event – as defined by authoritative data – have a higher probability of being relevant.

Table 1: Analysis of tweets for 2014 Southern England floods: comparison between number and relevance

Localization	**No. of tweets**	**Relevant**	**Percentage of relevant tweets**
Georeferenced	4	4	100%
Using local context	26	16	62%
Using global context	15	13	87%

We performed an analysis of the geolocated images using image processing techniques. Out of 348 images with geolocation information, 95 were found to be duplicates (ranging from 2 to 9 copies of individual images). We used Google Cloud Vision for image interpretation and found 73 unique tags, with the number of occurrences ranging from 1 (e.g., arch, aerial photography), to 23 (e.g. map), to 63 (e.g. water). This preliminary analysis provides an idea of the heterogeneity of the information available and of the potential benefits of complementing our analyses with image processing techniques such as a post-processing step. Future research will explore how to best integrate these two types of analysis (geolocation and image processing).

5 Exploring connected media

Even if we are able to increase the number of geolocated images from tweets from 5% to 16%, the number is still rather low for the purposes of a good coverage of the maps in the areas of interest.

We are therefore working on an explorative approach (as proposed in Di Blas et al., 2017) to identify further sources of information in social media. This exploratory approach allows us to find further social media content from other sources, such as flickr, YouTube and Instagram, and to refine the geolocation obtained by analysing tweets automatically.

An example is shown in Figure 3, a still from a video retrieved by following a link inserted in a tweet. The video posted on YouTube contains further geographical information which can be useful in determining its precise localization. The tweet itself specifies only 'Somerset' as a location, while the YouTube video itself adds more precision, mentioning 'Somerset Levels'.

Using this approach, in our case study we were able to identify 111 additional images or videos, 60% of which were relevant to the rapid mapping of the extent of the flooding.

Where Facebook was concerned, most of the posts were unavailable (either no longer available, or private) or did not contain media. In addition, very few of the retrieved posts were useful for rapid mapping.

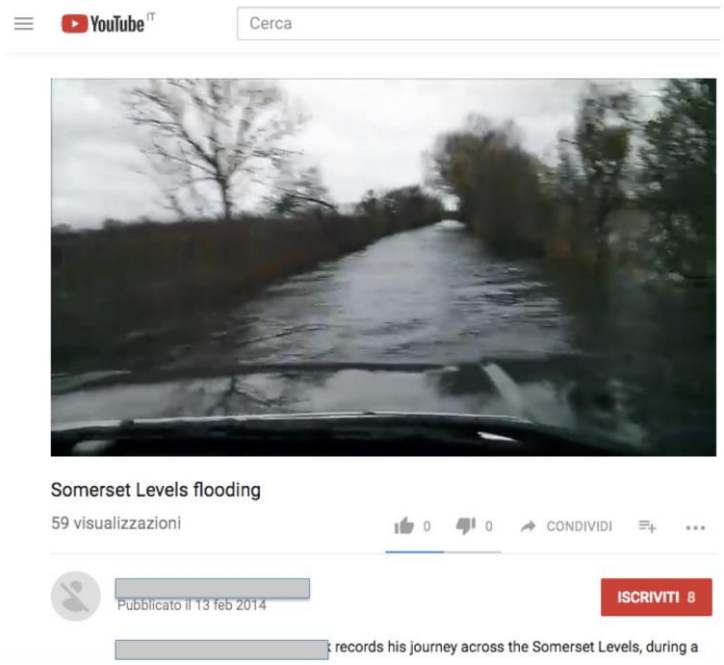

Figure 3: Linked YouTube video

6 Concluding remarks

In this paper, we have presented a context-based, exploratory approach to extracting geolocated images from tweets which refer to locations. The approach was tested in several case studies that looked at emergency management contexts, in particular floods, storms and earthquakes within the E2mC project (Resch et al., 2017).

In future research, we are planning to extend the exploratory approach, creating adaptive context-based crawlers to explore related posts, exploiting, for instance, collections like galleries in Flickr, or posts by the same authors in the same period on Twitter.

Acknowledgments

This work has been partially funded by the European Commission H2020 project E2mC - Evolution of Emergency Copernicus services, under project No. 730082. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

References

C. Castillo, C. (2016). *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

de Albuquerque, J.P, Herfort, B., Brenning, A., and Zipf, A. (2015). *A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management*. International Journal of Geographical Information Science, 29(4):667–689.

Di Blas, N.,Mazuran, M., Paolini, P., Quintarelli, E., Tanca, L. (2017). *Exploratory computing: a comprehensive approach to data sensemaking*. International Journal of Data Science and Analytics 3(1), 61-77.

Francalanci, C., Pernici, B., and Scalia, G. (2018) Exploratory Spatio-Temporal Queries in Evolving Information, in MATES 2017 Mobility Analytics for Spatio-temporal and Social Data Workshop @ VLDB. Lecture Notes in Computer Science, Vol. 10731. Springer, Cham.

Haklay, M.M. and Weber, P. (2008). *OpenStreetMap: User-generated street maps*. IEEE Pervasive Computing, 7(4):12–18.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). *Processing social media messages in mass emergency: A survey*. ACM Comput. Surv., 47(4):67, 1-38.

Jurgens, D. et al. (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. Proc. ICWSM 15, 188-197.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA, System Demonstrations, 55–60.

Middleton, S.E., Middleton, L., and Modafferi, S. (2014). *Real-time crisis mapping of natural disasters using social media*. IEEE Intelligent Systems, 29(2):9–17.

Paraskevopoulos, P. and Palpanas, T. (2016). *Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets*. Social Netw. Analys. Mining, 6(1):89:1–89:16.

Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J.L., Van Achte, T., Zeug, G., Mondardini, R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., and Rüping, S. (2017). E2mC: Improving Emergency Management Service Practice through Social Media and Crowdsourcing Analysis in Near Real Time, Sensors, 17(12).

Resch, B., Uslaender, F., and Havas, C. (2018). *Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment*. Cartography and Geographic Information Science, 45(4), 2018.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2013). *Tweet analysis for real-time event detection and earthquake reporting system development*. IEEE TKDE, 25(4):919–931, 2013.

Scalia, G. (2017). Network-based content geolocation on social media for emergency management, Master Thesis, Politecnico di Milano, Milano, Italy.