Extracting Fine-grained Implicit Georeferencing Information from Microblogs Exploiting Crowdsourced Gazetteers and Social Interactions

Laura Di Rocco
University of Genova
Genova, Italy
laura.dirocco@dibris.unige.it

Michela Bertolotto
University College Dublin
Dublin, Ireland
michela.bertolotto@ucd.ie

Barbara Catania
University of Genova
Genova, Italy
barbara.catania@unige.it

Giovanna Guerrini
University of Genova
Genova, Italy
giovanna.guerrini@unige.it

Tiziano Cosso
University of Genova
Genova, Italy
tiziano.cosso@gter.it

Abstract

The multifaceted nature of data generated by social media, along with its geographic component, can be exploited to better understand social dynamics and propagation of information. This short paper presents an ongoing project focused on the extraction of fine-grained geospatial knowledge and georeferencing information from social media activities, in terms of both content and interactions. We aim at investigating the possibility of geolocalising non-geotagged tweets (i.e., tweets without coordinate field) at a fine-level of detail, using ontologies in order to exploit semantics for improving geolocation quality. Geolocation can be further refined by taking social interactions among users into account.
Keywords: Microblogs, Geographic Information Retrieval, Crowdsourced Information, Fine-grained localization

1 Introduction

Consistent user-generated data represent a valuable source for the extraction of new types of information patterns and knowledge. The multifaceted nature of user-generated data, along with its geographic component, is being exploited to better understand social dynamics and propagation of information. Social media activities can be associated with both an explicit and an implicit geographic information component. Consider, for instance, Twitter as a typical example. In this case, georeferencing information can be explicitly available as metadata, such as the user profile location and the GPS coordinates of the device from which the activity is performed. By contrast, implicit georeferencing information can be inferred, with variable degree of confidence, by the message content itself, which may contain images, names of entities with known spatial location, or by the social relationships and interactions among users.

Our focus is on inferring the tweeting location (i.e., the position of the user when the tweet was sent) rather than the user home location. Georeferencing a tweet is useful for several applications. For instance, to create heat maps to highlight areas from which tweets are generated or areas which tweets refer to. Location inference on Twitter is a good way for detecting the outbreaks of disease and natural disaster. This could be very useful in applications such as flash mob, short term events or emergency response. We notice that only a small percentage of tweets is explicitly georeferenced, as location services of mobile devices are often disabled or switched off to save battery. Hence, considering implicit geospatial information allows an improvement of the resulting quality of the georeferencing process, in terms of completeness.

Users play an important role as information producers also for what concerns geospatial information itself, in Volunteered Geographic Information (VGI). Crowdsourced geospatial data is becoming very popular mainly due to its free availability and its constant updating. Among all projects for spatial data crowdsourcing, OpenStreetMap (OSM)[1] is by far the most popular and is characterized by information at a very fine level of detail. This crowdsourced information is not only rich from the spatial viewpoint but it is also associated with textual descriptions of geographical entities, that can thus be correlated with the references to such entities in the text of tweets. Specifically, OSM contains information about “local” geographic entities (and corresponding terms) that we cannot find in other GeoDBs (e.g., OSM contain information about vernacular names of places, that is, the name commonly used by local users to refer to a place). This is because OSM is enriched by the contribution of individuals who typically have very detailed local knowledge. This huge amount of knowledge is thus of great value in terms of completeness and coverage (both in width and depth), even if it may suffer from heterogeneity and accuracy issues.

These observations on Twitter and crowdsourced geographic databases support the idea of our project. With the aim of fully exploiting the (explicit and implicit) fine-grained georeferencing information made available by social media, the project relies on semantically enhanced and refined crowdsourced geospatial data to extract fine-grained implicit geoinformation contained in tweet contents. This geoinformation is further refined relying on social interactions among tweeting users.

The remainder of the paper is organized as follows.
Section 2 provides an overview of the approach and highlights its novelty with respect to the literature. Section 3 contains some details about the various steps we follow. Section 4 discusses our evaluation plan and concludes.

2 Overview and Novelty of the Approach

1https://www.openstreetmap.org

Overview. Our approach is graphically illustrated in Figure 1. Specifically, our georeferencing system first gathers, through the use of the Twitter streaming API, both explicitly geotagged tweets and tweets missing an explicit geotagging. The tweets stream comes from a specific geographical area of interest (e.g., specific areas within a city). We consider an area of interest A and a bounding box BB that includes the target area of interest A.

As a first, preliminary, and offline step we identify the set of keywords to be looked for in the tweets contents, relying on a given gazetteer (OSM in our case). This process is referred to as geoname extraction. From the area of interest A, we obtain a set of keywords K_A , corresponding to geonames describing georeferenced entities contained in area A, i.e., the textual descriptions of objects contained in this area. This set of keywords is extracted from (semantically enriched) OSM. The semantic enrichment (further discussed in Sec. 3) allows us to obtain new geographical knowledge that helps us improve the set of extracted geonames (and thus, ultimately, the quality of geotagging).

The set of tweets we consider in the online processing (extracted by using a filter function of Twitter Streaming API) then consists of: (i) explicit georeferenced tweets from area A (i.e., tweets associated with geographic coordinates contained in BB); and (ii) non georeferenced tweets containing keywords in K_A , i.e., geonames related to area A extracted in the initial step.

A filtering is then applied to assess how strongly the mention of a local entity is an indication of the tweet being written from that location (see Sec. 3). Social relationships among users and their activities (such as mentions and retweets) are then exploited to further refine tweet geopositioning, taking into account only the interactions that likely denote spatial proximity. Georeferencing information belonging to content and social interaction analysis, appropriately weighted according to the respective confidence, are finally merged.

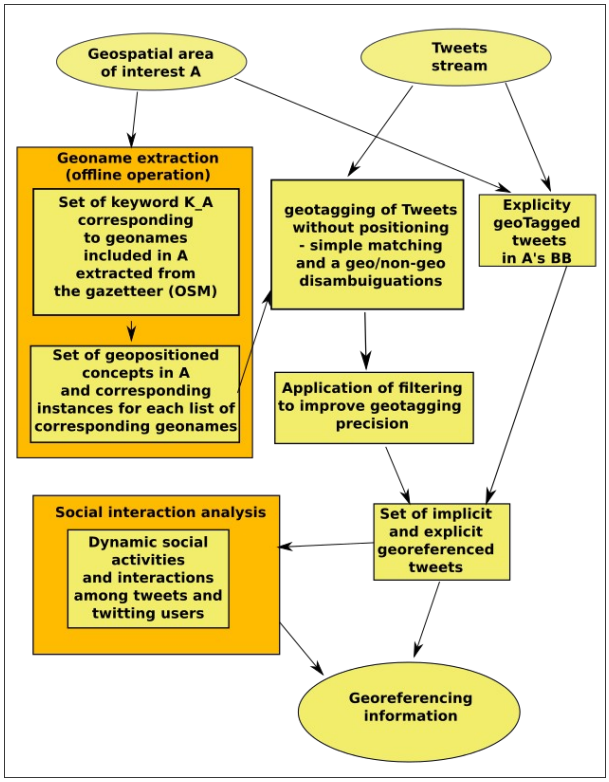
State of the Art. The problem of geotagging microblog messages has been largely investigated in the past. Some approaches (e.g., [4, 7]) extract georeferenced information using different techniques; however, the extracted georeferenced information typically refers to the detail level of city. In our work we want to infer georeferenced information at a finer level of detail, to differentiate tweets coming from different areas within the same city.

There are approaches in the literature [12] that separately use geolocation implicit in the text (message content) [3, 4] and in social activities (interaction) [2,6]. Our project deviates from these proposals in that it proposes the joint use of such information and the management of explicit geospatial information (and the extraction of implicit georeferencing information) at different levels of spatial resolution, where the most refined level corresponds to a higher detail than typically achieved by existing approaches.

Specifically, we want to achieve fine-grained georeferencing from microblog messages. In the literature, some research projects have tackled this issue. Gelernter et al. [5] improve the level of detail in geotagging analysing locations that occur in disaster-related social messages. Using a dataset containing messages exchanged during the Haiti earthquake of 2010 and Japan tsunami of 2011, they improve the location identification at the level of neighborhood, street, or building. Paraskevopoulos et al. [8] improve the geolocalisation based on the content similarities of tweets, as well as their time-evolution characteristics.

Differently from [5], we want to separate geolocalisation from a particular event. The approach presented in [8] could be the starting point of our investigation in order to improve geolocalisation using social interactions. However, in our work we propose to improve the precision of geotagging relying on a semantic enrichment of the detailed spatial data source used for georeferencing (in our case, OSM) by means of ontologies. Specifically, we want to localize both neighborhoods and points of interest within a city. Furthermore, we merge the capability to extract implicit geographic information from microblog texts with social interactions. For instance, the fact that a tweet *t* is a retweet or contains a mention of a user posting a tweet which is explicitly georeferenced can strengthen or weaken the confidence of the position inferred for the tweet *t*.

Figure 1: Overview of the approach



3 Inferring Geopositioning from Content & Social Activities

In this section we discuss the different steps that we follow in order to infer geolocation from microblog content and social activities.

Geonames Extraction through Crowdsourced Semantically Enhanced Gazetteers. Geopositions can be implicit in tweets that mention geographic entities. The names of these entities can be detected by using a Gazetteer. However, classical gazetteers, like GeoNames[2] , are not very

2 http://www.geonames.org/

useful as they are too coarse-grained. Therefore we decided to use OSM. Indeed, as already pointed out, we need a semantic gazetteer providing a high level of detail. In order to successfully exploit OSM for geoparsing, we consider OSM tags (e.g., name = * for tags related to a street or to a point of interest).

However, explicit geographic information in OSM (as in other VGI repositories) has inherent heterogeneity and quality issues, due to its crowdsourced nature. In particular, it has been noted that the semantic structure of OSM data is quite poor. Therefore our first step consists of semantically enhancing OSM data through the definition of a properly structured ontology that allows to classify OSM tags. OSM tags contain different types of information. For us, it is important to analyse the tags that contain geospatial names only.

As our objective is to semantify OSM tags, in order to define an ontology we start from the analysis of OSM tags employed in the context of a specific urban context: the city of London (UK). London was chosen because there is a big amount of OSM data referring to it, likely covering most of the relevant concepts in an urban context, and a constant flow of tweets geolocated in London. Thus it provides a good dataset. We then generalize the ontology concepts to encompass concepts that may occur in a generic city, trying to avoid, whenever possible, to focus on the specificities of a particular city.

Some researchers have tackled related issues. In particular, LinkedGeoData [1, 11] is a project that aims at linking OSM data to other LinkedData repositories (such as GeoNames and/or other online ontologies) by converting it to RDF so that it can be queried from a *SPARQL endpoint*. However, LinkedGeoData does not include all OSM entities and therefore it is not very useful for our purposes.

To this aim, we developed a facet ontology [9]. Facet ontologies classify objects using multiple taxonomies. A facet is a hierarchy of homogeneous concepts describing an aspect of the domain, where each term denotes an atomic concept. Each facet is designed separately, and models a distinct aspect of the domain. Each facet consist of a terminology, i.e., a finite set of names or terms, structured by a subsumption relation. In our ontology we defined facets corresponding to geophysical, geopolitical, and Point of Interest aspects. The use of facets takes into account the different aspects involved (i.e., natural area, political area and Point of Interest), thus obtaining a complete characterization of the domain of interest.

The developed ontology allows OSM data to be used as instances. This way we provide support for semantically searching OSM datasets. Conscious of the heterogeneous nature of geospatial data, we do not provide any contribution to the spatial component of the data, already well structured. Instead we aim at improving the non-spatial (semantic) content which is per se heterogeneous and only semi-structured. The non-spatial content is now accessible through a semantic structure.

Geotagging. The Twitter Streaming API is used to extract tweets with geonames obtained from our gazetteer. This integration works like a simple classifier. That is, we only look for occurrences of the geonames in the message text. This choice is due to the consciousness that microblogs messages are not always written in natural language.

Filtering. The association between a tweet and the position of the geonames mentioned in the text might not be accurate as such an association is only based on the assumption that the user writes about a place where she is which seems reasonable at least in a social network based on realtime writing and reading. Some filters are therefore applied to weigh the confidence of the tweet/location association by considering different characteristics and aspects of the tweet itself, e.g., the presence of an image, the device from which the tweet is written, and of the tweeting user, e.g., the user typology. More precisely, we filter out tweets that were not sent from a mobile device. We further assume that a tweet with images likely contains an image taken by the camera of the mobile device from which the tweet is written. For what concerns the users, we remove all tweets from users whose ratio of followers/following is higher than a given threshold value assuming that they would likely be famous people. The retweet of one of their tweets or their mentioning is typically not an indicator of the tweet being written in a location close to them. The more influential a user is, indeed, the more likely the users interacting with her would be from around the world.

Social interactions. Another source of implicit geopositioning in social media is related to social interactions among users. Since we are interested in tweeting locations rather that in user home locations, the social relationships most fruitfully exploited for refining the geopositioning are dynamic social interactions (i.e., retweeting, retweeting of the same tweet, mentions) rather than more stable relationships such as long-lasting following-follower relationships. For example, the retweeting of a geopositioned tweet can indeed be an indicator for positioning the retweeting user. Social relationships, by contast, have been mostly exploited in the literature for inferring user home location. It has been highlighted that, in the use of Twitter, usually more than 50% of accounts establish a relationship with other users living or stationed in the same place [10].

4 Preliminary Evaluation Plan & Conclusions

Although a comprehensive assessment of the validity of our approach has not been carried out yet, we have a clear plan for a two-fold evaluation. More specifically, we will perform the following two types of evaluation:

1. comparison between manually geotagged tweets and automatically geotagged tweets;
2. for geotagged tweets, comparison between the known (exact) position and the location inferred by our approach (only in case in which we have tweets with coordinates and containing geonames).

The first aim of our evaluation is to understand if our approach works correctly on a specific dataset. The dataset we selected contains English tweets from London (UK) and was extracted through Twitter Streaming API and contains approximately 100,000 tweets exchanged over a period of 4 hours.

Tweets are split in two categories:

1. G : set of tweets with coordinates (14497 tweets);
2. U : set of tweets geotagged without coordinates (81366 tweets).

We noticed that the percentage of tweets in G is relatively small. The applied filtering significantly reduced the cardinality of the two sets. This indicates that many people tweet from non-mobile devices and/or they have a lot of followers as compared to the number they are following. Furthermore, the percentage of non-geotagged tweets is relatively small. This is true in general. Indeed, after analysing an Italian dataset, we obtained proportionally the same results.

Since this is a preliminary evaluation, we only extracted tweets in which geonames keywords are neighborhoods of London. Initial experiments with a small subset of tweets have given encouraging indications on the viability of the proposed method.

The approach presented in this short paper brings a number of novel perspectives on inferring tweeting locations at a fine level of details. The ongoing experimental evaluation is aimed at demonstrating the benefits of the approach in terms of coverage and accuracy of the inferred location and its feasibility from a performance/scalability viewpoint.

References

[1] Auer, Sören, Jens Lehmann, and Sebastian Hellmann. *Linkedgeodata: Adding a spatial dimension to the web of data*. Springer Berlin Heidelberg, 2009.

[2] Backstrom, Lars, Eric Sun, and Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity."*Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[3] Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond trending topics: Real-world event identification on twitter." (2011).

[4] Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010.

[5] Gelernter, Judith, and Nikolai Mushegian. "Geo-parsing Messages from Microtext." *Transactions in GIS* 15.6 (2011): 753-773.

[6] Kamath, Krishna Yeshwanth, and James Caverlee. "Transient crowd discovery on the real-time social web." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011.

[7] Kinsella, Sheila, Vanessa Murdock, and Neil O'Hare. "I'm eating a sandwich in Glasgow: modeling locations with tweets." *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011.

[8] Paraskevopoulos, Pavlos, and Themis Palpanas. "Fine-Grained Geolocalisation of Non-Geotagged Tweets." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015.

[9] Ranganathan, Shiyali Ramamrita. "Prolegomena to library classification." *The Five Laws of Library Science* (1967).

[10] Sadilek, Adam, Henry Kautz, and Jeffrey P. Bigham. "Finding your friends and following them to where you are." *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012.

[11] Stadler, Claus, et al. "Linkedgeodata: A core for a web of spatial open data."*Semantic Web* 3.4 (2012): 333-354.

[12] Stefanidis, Anthony, Andrew Crooks, and Jacek Radzikowski. "Harvesting ambient geospatial information from social media feeds." *GeoJournal* 78.2 (2013): 319-338.