Thesis Proposal:

Semantically enriched methods for next generation microblog message fine-grained geolocalization

Laura Di Rocco

Oct, 2016

1 Motivation

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. Since explicit tagging is used only in a small percentage of tweets, we will use geospatial information implicit in the messages to improve the resolution of the georeferencing process. 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 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). Goodchild [21] defines VGI as "[...] *a special case of the more general Web phenomenon of user-generated content*[...]". 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 a spatial viewpoint but 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.

Therefore, we have chosen Twitter as a specific example of microblog social media to determine a model of explicit and implicit geospatial information. The particular model of implicit geospatial information on Twitter allows us to use our ontology to find a set of tweets sufficiently accurate. This set will be used as domain for definition of algorithms that will extract a complete and accurate geospatial reference, e.g., algorithms about social relationship among users.

We focus on OpenStreetMap dataset and Twitter dataset only to better understand the methodology. But it is possible to apply these techniques to other geospatial dataset (for explicit geospatial information) and other forms of microblog (for implicit geospatial information).

The document is structured as follows: Section 2 introduces the background related to the field of this proposal, Section 3 analyzes the state of the art, Section 4 provides detail about research plan and current status and in Section 5 we explain the future work.

2 Background: Social, Geospatial, Crowdsourced data

In recent years, social networks achieved an important role in social life. People use social networks to share a considerable and heterogeneous quantity of information. The resulting large amount of data is increasingly being used for analysis of social behaviour.

Network generated data is mainly unstructured. It includes big quantities of metadata on the web that grows exponentially, from which we can extract structured information with semantic analysis techniques. Specifically, social media data is a form of network generated data. Among social media data, Microblogs are very common (e.g. Twitter). Microblogging is a broadcast medium that exists in the form of blogging. However, a microblog differs from a traditional blog in that its content is typically smaller in both actual and aggregated file size. Microblogs contain a lot of abbreviations, misspelling, etc. that need to be taken into account. The most famous microblog is Twitter.

Twitter posts may be associated with an explicit geospatial information in form of GPS coordinate, and an implicit geospatial information in form of the text message.

Data generated by social media is a form of Big Data, a term that refers to very large amounts of data, ranging from a few dozen terabytes to many petabytes. While big data

[1] `https://www.openstreetmap.org`

can have various properties, such velocity, volume and variety [41, 42, 33, 35], we focus on the variety of data. Indeed big data, as a variety, can be produced by various different domains. The most important features of social media data (and also crucial for our research interests) are:

- streamed (Velocity),
- semistructured (Text).

Another important type of Big Data is geospatial data. Geospatial data have two components: a spatial component which specifies their location in space, and a non-spatial component which specifies other characteristics. For instance, for a road these could include the road name, the road type, the company in charge of road maintenance, the traffic volume, etc. These two types of information (spatial and non-spatial) are called the geometry and the attributes of the spatial datum under examination, respectively.

Geospatial data can also be characterised in terms of their topology. Geometry and topology are properties specific to spatial data. Geometry is concerned with shape (point, line, or region), extension (coordinates defining the point, line, region), position (in a geographic reference system). Topology is concerned with the connectivity between spatial entities, independently of their coordinates (e.g., two entities intersect, are disjoint, adjacent, overlapping, ...). Topology provides qualitative rather than quantitative information (e.g., if two entities intersect, the shape, extension, and position of the intersection does not matter; if they are disjoint, the distance separating them does not matter, etc.). Topological relations are invariant under continuous transformations (e.g., we can deform the shape of the two entities while preserving the fact that they intersect, or they are disjoint).

The characteristics of geospatial data and the different sources generating it, create an interesting domain for Big Data. As in many other fields, different types of geospatial data exist, namely authoritative datasets (e.g. GeoNames[2]) and crowdsourced datasets (e.g. OSM[3]).

In particular, crowdsourced geospatial data are becoming very popular mainly due to its free availability and its constant updating. Among all projects for spatial data crowdsourcing, OSM is by far the most popular. OSM is a collaborative project to create a free editable map of the world via crowdsourced data [4].

OSM is a considerable source of Big geospatial Data. Indeed, each day users add up to three millions nodes [5].

In particular, in our research, we focus on a specific type of geospatial data, i.e., crowdsourced geospatial data also called Volunteered Geographic Information (VGI). Goodchild [21] defines VGI as "[...] *a special case of the more general Web phenomenon of user-generated content*[...]". This is because it is gathered by individuals with no formal training. The quality and reliability of VGI is a matter of much debate and an active research topic. For instance, questions around the quality of the data and the credibility of the volunteers [16, 23]. Volunteers use tags to annotate geospatial object on VGI. The simplicity to use tags allows users to put a lot of tags on the objects. This creates a big and noisy tagspace in which it is harder to find material tagging by other people. This is due also to psychological aspects of tagging: people think differently and so use tagging differently[10]. In OSM, this means that not all the objects contain the correct tags, that some objects may contain the same tags though they are of different types, and that some objects of

[2] `http://www.geonames.org/`
[3] `https://www.openstreetmap.org/`
[4] `http://wiki.openstreetmap.org`
[5] `http://wiki.openstreetmap.org/wiki/Stats`

the same type may contain different tags. In another example, Koukoletsos et al.[32] propose an automated feature-based matching method for VGI. They develop a method to match the geometry of a VGI dataset to a reference dataset, to analyse the quality of data. Lamprianidis et al.[34] propose a method to integrate crowdsourced Point of Interest (PoI) data from multiple heterogeneous web sources.

3 State of the Art

In this section, we discuss related work on location inference from microblog messages, relying on semantic geographic gazetteers. There are different techniques to extract implicit geographic information from social media messages. The discussion will investigate the difference among the different approaches. We also highlight that the novelty of our approach is due to the combination of existing techniques in order to improve completeness and correctness of geotagging. In the survey [1], a general overview of different methods for location inference can be found. We start with an introduction of the problem of geotagging in general and after that we explain in detail the difference in geotagging level that we can achieve.

3.1 Geotagging

The problem of geotagging microblog messages has been largely investigated in the past. Some approaches (e.g., [12, 31]) extract georeferenced information using different techniques; however, the extracted georeferenced information typically refers to the detail level of city. In our work we infer georeferenced information at a finer level of detail, to differentiate tweets coming from different areas within the same city.

Data are said to be implicitly (or indirectly) georeferenced [26] when they are not associated with explicit geospatial references (such as positioning on maps or spatial coordinates), rather they are referenced by place names, geocodes and addresses. The terms highlights the fact that additional steps are required to identify the locations on maps.

Georeferencing by place name is the most common form of referencing a geographic location and is an informal means of georeferencing. We use place names in conversations, correspondence, reporting, and documentation. Dictionaries of placenames are called gazetteers [22]. They contain descriptive information about named places, which can include their geographic locations, types/categories, and other infomation.

Implicit georeferencing information has been exploited, for instance, for localizing news on maps [59].

There are approaches in the literature [57] that separately exploit geolocation information implicit in the text (message content) [9, 12] and in social activities (interactions) [4, 30, 52]. 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.

Geotagging consists of toponym recognition and toponym resolution. We first provide an overview for what concerns the former. The latter, instead, is eliminating the geo/non-geo ambiguity (e.g., Washington can be a city in the USA or the name of a person). There are different strategies to do this:

1. finding names in the text that exist in a gazetteer[40, 39, 53, 2, 24];
2. using Name Entity Recognition techniques[49, 58];

3. using a geographic ontology in order to understand the context of the text[58, 49].

3.1.1 City-level location

In general, in Geographical Information Retrieval, is important understand what does it means geolocalization at city-level. We can subdivide the city-level localization problem in two different problems:

1. home localization,
2. twitting localization.

Home localization is the problem related to the localization of the user. Instead, twitting localization is the problem related to the localization of Twitter messages.

The work of Cheng et al.[12] uses a probabilistic framework to estimate the city-level coordinates of tweets based on the text message. They do not use geotags of a single tweet but use the place information field extracted from the user profile. They identify words that have local focus and then model their geographic distribution, i.e., they build a statistical predictive model. A similar approach can be found in [31]. They model locations from zip code to country level using a probability distribution of terms associated with a location. Chandra et al. [11], use a probabilistic framework to estimate the city-level Twitter user location. The probabilities are based on the contents of the tweet messages with the aid of reply-tweet messages generated from the interaction between different users in the Twitter social network. Ikawa et al. [27] want to determine the location where the messages are generated without using geotagging but only relying on the message content. To do this, they use a training set and then they estimate location with basic machine learning algorithms.

Mahamud et al. [43] use classifiers on three different terms: words, hashtags, and place names. These classifiers can be created for any level of granularity for which they have ground truth.

Cunha et al.[14] analyse tweets to find a spatio-temporal pattern. They use data mining to analyse different types of information from tweets: spatial, temporal, social and content. Another example of a system that studies the interaction between Twitter and VGI is GeoSEn [46].

Finally, Schulz et al. [54, 55] propose a method that relies on a multi-indicator spatial approach to solve the problem of disambiguation of toponyms. The result is an algorithm for determining the location where a tweet was generated and also the user's home.

3.1.2 Fine-grained location

In the literature, some research projects have tackled the issue of fine-grained georeferencing from microblog messages. Gelernter et al. [18, 19] 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. [48] improve the geolocalisation based on the content similarities of tweets, as well as their time-evolution characteristics.

Differently from [18], we want to separate geolocalisation from a particular event. The approach presented in [48] could be the starting point of our investigation in order to improve geolocalisation using social interactions. However, in our work we propose to

Table 1: Overview of different approaches in geotagging

	City-level location			Fine-grained location	
	Gazetteers	NER	Ontologies	Similarity of texts	Event-detection
Amitay et al.[2]	✓				
Lieberman, Samet, and Sankaranarayanan[40]	✓				
Lieberman, Samet, and Sankaranarayanan[39]	✓				
Grover et al.[24]	✓				
Purves et al.[49]		✓	✓		
Stokes et al.[58]		✓	✓		
Cheng, Caverlee, and Lee[12]	✓	✓(ML techniques)			
Kinsella, Murdock, and O'Hare[31]	✓	✓(ML techniques)			
Chandra, Khan, and Muhaya[11]	✓(ML techniques)				
Ikawa, Enoki, and Tatsubori[27]		✓(Context of messages)			
Mahmud, Nichols, and Drews[43]		✓			
Cunha, Soares, and Rodrigues[14]	✓	✓(spatio-temporal patterns)			
Schulz et al.[54]	✓	✓(multi-indicator spatial approach)			
Gelernter and Mushegian[18]					✓
Paraskevopoulos and Palpanas[48]				✓	

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 want to refine the implicit geographic information extracted from microblog texts by exploiting information associated 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 t. Other approaches for inferring fine-grained location of messages include [37, 38]. However, they focus mainly on Points of Interest (and in some cases district) of a particular place.

In Table 1, we have a summary of the state of the art.

3.2 Geospatial Ontologies

To better understand the role of a semantic enrichment in our thesis proposal, we talk about geospatial ontology in general. We analyse here, the state of the art related to classification of geospatial data.

An ontology is defined as: *"a specification of a conceptualization."* [25]. Ontologies have been used for a set of tasks: improving communication between agents (human or software), reusing data models, developing knowledge schemas, etc. . All these tasks deal with interoperability issues and can be applied in different domains.

Ontologies of the geographic world are important to allow the sharing of geographic data among different communities of users. A geo-ontology provides a description of geographical entities.

A lot of geospatial ontologies have been defined and can be used by different applications. We first discuss some common ontologies that we analyse for our work, then we present specific approaches related to ontologies for VGI.

To position our work in the context of GIS, we analyse the follows ontologies:

- GeoNames;
- Qall-me;
- The administrative geographic and civil voting area ontology;
- Dbpedia.

GeoNames is a geographical database available and accessible through various web services. The GeoNames database contains over 10.000.000 geographical names corresponding to over 7.500.000 unique features. All features are categorized into one of nine feature

classes and further subcategorized into one of 645 feature codes. Each GeoNames feature is represented as a web resource identified by a stable URI. This URI provides access, through content negotiation, either to the HTML wiki page, or to a RDF description of the feature, using elements of GeoNames ontology. This ontology describes the GeoNames features properties using the web ontology language, the feature classes and codes being described in the SKOS language. GeoNames consists of various locations of all countries. It includes as geographical data: place names in various languages, latitude, longitude[6], altitude and class.

Qall-me ontology[7] is a domain-specific ontology that has been developed and applied for question answering in the domain of tourism[47].

"The administrative geography and civil voting area ontology" [8] is an ontology developed by Ordnance Survey[9]. It is an ontology describing the administrative and voting area geography of Great Britain. This is related to Great Britain but it contains class `Borough` without other subclasses because, for this domain, it is not important to have a high level of detail.

DBpedia[10] is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. The goal is to make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might be inspired new mechanisms for navigating, linking, and improving the encyclopedia itself. The DBpedia Ontology is a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia. The ontology currently covers 685 classes which form a subsumption hierarchy and are described by 2,795 different properties.

Several other researchers present techniques to create and organize ontologies related to VGI. Beard [8] introduces an ontology based gazetteer model for organizing VGI. His aim is to try to associate VGI contribution to place such that overtime places may be characterized through these contributions. Stadler et al.[56] presented an ontology called LinkedGeoData that links OSM information to DBpedia, GeoNames and others ontologies.

In particular, LinkedGeoData [3, 56] 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 that we need and therefore it is not very useful for our purposes.

Finally, Fonseca et al. [17] present an ontology to classify geographic elements with respect to not only their geometrical features, but also their attribute values, i.e. their semantic features. In this work they integrate vector-based GIS (geometrical feature) that imports raster data (attribute values) or than raster-based GIS than imports vector data. The result of this classification with an ontology support is a set of images indexed not only by its semantics but also by its attribute values. In this way, they obtain not only a static polygon but also a set of semantic feature and its corresponding values. This type of approach is very useful in case you want analyse and manage observation on continues spatio-temporal data.

In the geospatial domain, there are a lot of approaches that try to use ontologies to semantifying these type of data[28, 29]. We have analysed research about OSM. OSM

[6] according to the WGS84

[7] EU-funded project - `http://qallme.fbk.eu/index.php?location=ontology`

[8] `http://data.ordnancesurvey.co.uk/ontology/admingeo/`

[9] `https://www.ordnancesurvey.co.uk/`

[10] `http://dbpedia.org/`

offers an open and easy to use platform that enables contributors to upload geographic information collected from mobile devices or aerial images. There is no formal ontology or vocabulary of predefined tags that have to be adopted by the users. For this reason, there exists a lot of work about semantifying OSM. For instance, Baglatzi et al. [5] shows a way to bridge the gap between ontological and crowdsourcing practices. To create this bridge they use ontological alignment approach. They align OSM tags to the DOLCE Ultralite top level ontologies [45].

There is also an important project called OSMOnto[11] [13]. This is an ontology for tags. The purpose of the ontology of tags is to stay as close as possible to the structure of the OSM files in order to facilitate database querying. This means that they do not try to correct any possible conceptual mistakes in the taxonomy of OSM tags, but rather have it reflected faithfully in the structure of the ontology.

Another example of semantic research work on OSM is the OpenStreetMap Semantic Network[12] [7, 6]. It is a Semantic Web resource extracted from the OSM Wiki website, encoded as a SKOS vocabulary [13]. It contains a machine-readable representation of OSM tags, and several semantic relationships among them.

Another problem that we have to cope with is language. When we search information about somenthing, we usually use our language. For instance, we are Italian, so we usually search information in Italian. But, if we search 'Londra' and not 'London', we would find less information in Italian then in English. To solve this kind of problems, there are several research works. For instance, Laurini [36] uses ontologies among gazetteers to try to create a bridge connecting the same concept in different languages.

4 An Approach to Georeferencing tweets: Overview and Current Status

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 allows us to obtain new geographical knowledge that helps us improve the set of extracted geonames[50, 20, 44] (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.

[11] It is possible to find more information here: `http://wiki.openstreetmap.org/wiki/OSMonto`

[12] It is possible to find more information here: `http://wiki.openstreetmap.org/wiki/OSM_Semantic_Network`

[13] Simple Knowledge Organization System - `http://www.w3.org/2004/02/skos/`

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.

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.

With this approach we will develop algorithms for geolocalising tweets at fine-grained detail. We will first develop an algorithm for toponyms extraction using a semantic support. The first steps will be:

TASK 1 extraction of toponyms from tweets, in order to improve the geographical position inference process;

TASK 2 integration of the approach proposed in [15] with the ability to associate geo-positions through the usage of toponyms.

In order to extract toponyms from text (TASK 1), we plan to consider two different approaches:

1. Improve matching by relying on a geospatial ontology. A semantic support can help us to find toponyms using also the context of the tweet (e.g. the presence of word "cinema" allow us to search a toponym in the proper class of the ontology), improving the precision of matching.

2. Rely on NLP classifiers in order to identify reference to geographical locations from texts in order to identify toponyms by the context of terms (prepositions, verbs, ...).

In TASK 2, we plan to integrate the approaches [48, 15] in order to:(i) increase the number of tweets with which a position can be associated; (ii) evaluate whether positions inferred through toponyms can be used to revise/refine positions inferred through the techniques presented in [48]. These two tasks will be completed during the exchange period in Universitè Paris Descartes.

Another important aspect that we want to consider is the analysis of geographic distances between places. More specifically, if we find a name we will be able to disambiguate if it is a geographical name or not using the distance between places.

The last step that we would investigate (a long time period work), will be the study of social relationship.

4.1 Current Status of Research

The approach presented in this proposal brings a number of novel perspectives in inferring tweeting locations at a fine level of detail.

We published, as a short paper, an extract of this project in AGILE 2016 Conference [15]. In this short paper, we have presented our 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 want to investigate the possibility of geolocalising non-geotagged tweets (i.e., tweets without coordinate field) at a high-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.

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.

The approach, which we propose, starts from a collaboration with Dr. Tiziano Cosso (Gter, spin-off Università di Genova) and Dr. Michela Bertolotto (University College Dublin). These collaborations can improve our knowledge in the GIS field due to their expertise.

In this first year, we have analyzed the current state of the art related to Geographical Information Retrieval. We can assert that, at the end of this first year, we have a good overview of the recent work in this field. However, we can not assert that the state of the art analysis is completely done. The next aim will be to analyze in detail the techniques that we want to use to design and develop the algorithms.

To improve our knowledge in Computer Science and in the Geographic Information Retrieval field, and also in general in Information Retrieval, during the first year, some courses and doctoral schools will be followed. Here, the relevant courses[14]:

- Algorithmic methods for mining large graphs, Prof. Aristides Gionis
- Tools and techniques for massive data analysis, Dr. Giuseppe Fiameni
- Semantic Web, Prof. Barbara Catania, Prof. Giovanna Guerrini, Prof. Viviana Mascardi
- Introduction to consultation and analysis of geographical data using GIS software, Prof. Bianca Federici

Finally, at the end of first year, we started a collaboration with Prof. Themis Palpanas, Universitè Paris Descartes, Paris, France, and his PhD Student, Pavlos Paraskevopoulos. Their research is concentrated on problems related to online and offline data analytics (focusing on fast streams and massive data collections). Their group has world-leading expertise in mining non-geotagged tweets.

The joint collaboration will improve my knowledge in social media data mining and information retrieval. Our aim is to use geographical knowledge, like semantic gazetteer, in order to improve geolocalization of Twitter messages.

5 Research Plan

The general objective of this thesis proposal is to find implicit geographical information in a microblog. To this aim, the project will rely on crowdsourced geographical data (Open Street Map).

The specific objectives of this thesis proposal are as follows:

- Investigation of a social networks to understand the difference between explicit and implicit geospatial information associated to social media data. Definition of a model to manage geospatial information extracted from social media data.
- Characterization of explicit geospatial information available in open sources format (OSM) and evaluation of their quality[51]. One solution for this characterization is the use of ontologies. In this way we can extract crowdsourced information that will be correctly classified, therefore improving accurateness and completeness.

[14]The other information related to courses, exams and schools are presented in my annual report

- Definition of algorithms for extracting implicit geospatial information from Twitter data. The extraction algorithms must be able to manipulate geo-spatial information at different levels of detail. Indeed, the link between well-formed explicit geospatial information (e.g., ontology) and implicit geospatial information is resolved by these algorithms.
- Analysis of the proposed algorithms in terms of quality of the extracted geospatial information (completeness, accuracy).
- Implementation of the proposed algorithms in a parallel context (MapReduce), in order to ensure scalability, possibly using NoSQL data management systems. This is most important in this project because the social networks' data are very huge. This help us also to improve efficiency.
- Performance evaluation of the proposed algorithms in some reference/sample scenarios.

To demonstrate the scheduling of our work, we show with Tables 2,3 how we plan to utilize time in order to allocate tasks towards our objectives for the next two years.

Table 2: Workplan

Activity	**Description**	**Months**	**Colour**
1	State of the art related to Geographic Information Retrieval	1st Year	
2	Development of the geospatial ontology	1st Year	
3	Definition of algorithms	6	
4	Development of algorithms	6	
5	Evaluations	6	
6	Development a parallel algorithm using NoSQL techniques	6	
7	Evaluations	6	
8	PhD thesis writing	4	

Table 3: **A** = period abroad

Activity	Year1												Year2												Year3											
1																																				
2																																				
3										**A**	**A**	**A**																								
4										**A**	**A**	**A**																								
5																																				
6																																				
7																																				
8																																				

This is a general aim of our PhD project.

References

[1] Oluwaseun Ajao, Jun Hong, and Weiru Liu. "A survey of location inference techniques on Twitter". In: *Journal of Information Science* 41.6 (2015), pp. 855–864.

[2] Einat Amitay et al. "Web-a-where: geotagging web content". In: *Proceedings of SIGIR '04 conference on Research and development in information retrieval* (2004), pp. 273–280. ISSN: 10980121. DOI: 10.1145/1008992.1009040. URL: http://portal.acm.org/citation.cfm?id=1009040.

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

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

[5] Alkyoni Baglatzi, Margarita Kokla, and Marinos Kavouras. "Semantifying OpenStreetMap". In: *Proceedings of the 5th International Terra Cognita Workshop*. Citeseer. 2012, pp. 39–50.

[6] Andrea Ballatore and Michela Bertolotto. "Semantically Enriching VGI in Support of Implicit Feedback Analysis". English. In: *Web and Wireless Geographical Information Systems*. Ed. by Katsumi Tanaka, Peter Fröhlich, and Kyoung-Sook Kim. Vol. 6574. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 78–93. ISBN: 978-3-642-19172-5. DOI: 10.1007/978-3-642-19173-2_8. URL: http://dx.doi.org/10.1007/978-3-642-19173-2_8.

[7] Andrea Ballatore, Michela Bertolotto, and DavidC. Wilson. "Geographic knowledge extraction and semantic similarity in OpenStreetMap". English. In: *Knowledge and Information Systems* 37.1 (2013), pp. 61–81. ISSN: 0219-1377. DOI: 10.1007/s10115-012-0571-0. URL: http://dx.doi.org/10.1007/s10115-012-0571-0.

[8] Kate Beard. "A semantic web based gazetteer model for VGI". In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. ACM. 2012, pp. 54–61.

[9] Hila Becker, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." In: *ICWSM* 11 (2011).

[10] Grigory Begelman, Philipp Keller, Frank Smadja, et al. "Automated tag clustering: Improving search and exploration in the tag space". In: *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*. 2006, pp. 15–33.

[11] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. "Estimating Twitter User Location Using Social Interactions–A Content Based Approach". In: *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* (2011), pp. 838–843. DOI: 10.1109/PASSAT/SocialCom.2011.120. URL: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6113226.

[12] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. "You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. Toronto, ON, Canada: ACM, 2010, pp. 759–768. ISBN: 978-1-4503-0099-5. DOI: 10.1145/1871437.1871535. URL: http://doi.acm.org/10.1145/1871437.1871535.

[13] Mihai Codescu et al. "Osmonto-an ontology of openstreetmap tags". In: *State of the map Europe (SOTM-EU) 2011* (2011).

[14] Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. "TweeProfiles: detection of spatio-temporal patterns on Twitter". In: *Advanced Data Mining and Applications*. Springer, 2014, pp. 123–136.

[15] Laura Di Rocco et al. “Extracting Fine-grained Implicit Georeferencing Information from Microblogs Exploiting Crowdsourced Gazetteers and Social Interactions”. In: *Proceedings of the 19th AGILE International Conference on Geographic Information Science*. 2016.

[16] Andrew J Flanagin and Miriam J Metzger. “The credibility of volunteered geographic information”. In: *GeoJournal* 72.3-4 (2008), pp. 137–148.

[17] Frederico T Fonseca et al. “Using ontologies for integrated geographic information systems”. In: *Transactions in GIS* 6.3 (2002), pp. 231–257.

[18] Judith Gelernter and Nikolai Mushegian. “Geo-parsing Messages from Microtext”. In: *Transactions in GIS* 15.6 (2011), pp. 753–773.

[19] Judith Gelernter et al. “Automatic gazetteer enrichment with user-geocoded data”. In: *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD ’13* (2013), pp. 87–94. DOI: 10.1145/2534732.2534736. URL: http://dl.acm.org/citation.cfm?doid=2534732.2534736.

[20] Fausto Giunchiglia et al. “A Facet-Based Methodology for the Construction of a Large-Scale Geospatial Ontology”. English. In: *Journal on Data Semantics* 1.1 (2012), pp. 57–73. ISSN: 1861-2032. DOI: 10.1007/s13740-012-0005-x. URL: http://dx.doi.org/10.1007/s13740-012-0005-x.

[21] Michael F Goodchild. “Citizens as sensors: the world of volunteered geography”. In: *GeoJournal* 69.4 (2007), pp. 211–221.

[22] Michael F Goodchild and Linda L Hill. “Introduction to digital gazetteer research”. In: *International Journal of Geographical Information Science* 22.10 (2008), pp. 1039–1044.

[23] Michael F Goodchild and Linna Li. “Assuring the quality of volunteered geographic information”. In: *Spatial statistics* 1 (2012), pp. 110–120.

[24] Claire Grover et al. “Use of the Edinburgh geoparser for georeferencing digitized historical collections”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368.1925 (2010), pp. 3875–3889. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2010.0149. URL: http://rsta.royalsocietypublishing.org/content/368/1925/3875.

[25] Thomas R Gruber. “A translation approach to portable ontology specifications”. In: *Knowledge acquisition* 5.2 (1993), pp. 199–220.

[26] Linda L Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.

[27] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. “Location inference using microblog messages”. In: *Proceedings of the 21st international conference companion on World Wide Web - WWW ’12 Companion*. New York, New York, USA: ACM Press, 2012, p. 687. ISBN: 9781450312301. DOI: 10.1145/2187980.2188181. URL: http://dl.acm.org/citation.cfm?doid=2187980.2188181.

[28] Krzysztof Janowicz and Carsten Keßler. “The role of ontology in improving gazetteer interaction”. In: *International Journal of Geographical Information Science* 22.10 (2008), pp. 1129–1157.

[29] Krzysztof Janowicz et al. “Similarity as a quality indicator in ontology engineering”. In: *Formal Ontology in Information Systems* (2008), p. 92.

[30] Krishna Yeshwanth Kamath and James Caverlee. “Transient crowd discovery on the real-time social web”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011.

[31] Sheila Kinsella, Vanessa Murdock, and Neil O’Hare. “"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*. SMUC ’11. Glasgow, Scotland, UK: ACM, 2011.

[32] Thomas Koukoletsos, Mordechai Haklay, and Claire Ellul. “Assessing data completeness of VGI through an automated matching procedure for linear data”. In: *Transactions in GIS* 16.4 (2012), pp. 477–498.

[33] Alexandros Labrinidis and HV Jagadish. “Challenges and opportunities with big data”. In: *Proceedings of the VLDB Endowment* 5.12 (2012), pp. 2032–2033.

[34] George Lamprianidis et al. “Extraction, integration and analysis of crowdsourced points of interest from multiple web sources”. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. ACM. 2014, pp. 16–23.

[35] Douglas Laney. “3D data management: Controlling data volume, velocity and variety”. In: *META Group Research Note* 6 (2001).

[36] Robert Laurini. “Geographic Ontologies, Gazetteers and Multilingualism”. In: *Future Internet* 7.1 (2015), pp. 1–23.

[37] Chenliang Li and Aixin Sun. “Fine-grained location extraction from tweets with temporal awareness”. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR ’14*. New York, New York, USA: ACM Press, 2014, pp. 43–52. ISBN: 9781450322577. DOI: 10.1145/2600428.2609582. URL: http://dl.acm.org/citation.cfm?id=2600428.2609582.

[38] Guoliang Li et al. “Effective location identification from microblogs”. In: *Proceedings - International Conference on Data Engineering* (2014), pp. 880–891. ISSN: 10844627. DOI: 10.1109/ICDE.2014.6816708. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6816708.

[39] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. “Geotagging with local Lexicons to build indexes for textually-specified spatial data”. In: *Proceedings - International Conference on Data Engineering* May (2010), pp. 201–212. ISSN: 10844627. DOI: 10.1109/ICDE.2010.5447903.

[40] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. “STEWARD: architecture of a spatio-textual search engine”. In: *Proceedings of the ...* c (2007), pp. 186–193. DOI: 10.1145/1341012.1341045. URL: http://portal.acm.org/citation.cfm?id=1341045.

[41] Steve Lohr. “The age of big data”. In: *New York Times* 11 (2012).

[42] Sam Madden. “From databases to big data”. In: *IEEE Internet Computing* 3 (2012), pp. 4–6.

[43] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. “Home Location Identification of Twitter Users”. In: *"ACM Transactions on Intelligent Systems and Technology"* 5.3 (2014), pp. 47–69. ISSN: 21576912. DOI: 10.1145/2528548. URL: http://arxiv.org/abs/1403.2345.

[44] Vincenzo Maltese and Feroz Farazi. *A semantic schema for GeoNames*. A significant variation of this report has been accepted to the conference INSPIRE 2013 (http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2013/). Trento, 2013. URL: http://eprints.biblio.unitn.it/4088/.

[45] Claudio Masolo et al. *Wonderweb deliverable d17. the wonderweb library of foundational ontologies and the dolce ontology*. Citeseer, 2002. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.4243.

[46] Maxwell Guimaraes de Oliveira et al. "Leveraging VGI for Gazetteer Enrichment: A Case Study for Geoparsing Twitter Messages". In: *Web and Wireless Geographical Information Systems*. Springer, 2015, pp. 20–36.

[47] Ozer Ozdikis, Fatih Orhan, and Furkan Danismaz. "Ontology-based recommendation for points of interest retrieved from multiple data sources". In: *Proceedings of the International Workshop on Semantic Web Information Management*. ACM. 2011, p. 1.

[48] Pavlos Paraskevopoulos and Themis Palpanas. "Fine-Grained Geolocalisation of Non-Geotagged Tweets". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ASONAM '15. Paris, France: ACM, 2015.

[49] Ross S. Purves et al. "The Design and Implementation of SPIRIT: A Spatially Aware Search Engine for Information Retrieval on the Internet". In: *Int. J. Geogr. Inf. Sci.* 21.7 (Jan. 2007), pp. 717–745. ISSN: 1365-8816. DOI: 10.1080/13658810601169840. URL: http://dx.doi.org/10.1080/13658810601169840.

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

[51] Theodoros Rekatsinas et al. "Finding Quality in Quantity: The Challenge of Discovering Valuable Sources for Integration". In: CIDR. 2015.

[52] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. "Finding Your Friends and Following Them to Where You Are". In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: ACM, 2012.

[53] Hanan Samet et al. "Reading news with maps by exploiting spatial synonyms". In: *Communications of the ACM* 57.10 (2014), pp. 64–77. ISSN: 00010782. DOI: 10.1145/2629572. URL: http://dl.acm.org/ft{_}gateway.cfm?id=2629572{\&}type=html.

[54] Axel Schulz et al. "A Multi-Indicator Approach for Geolocalization of Tweets". In: *Seventh International AAAI Conference on Weblogs and Social Media* (2013), pp. 573–582. DOI: papers3://publication/uuid/62449928-74D1-4674-A1A7-24D5F6813F85.

[55] Axel Schulz et al. "Evaluating Multi-label Classification of Incident-related Tweets". In: *Making Sense of Microposts (# Microposts2014)* (2014), p. 7.

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

[57] Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. "Harvesting ambient geospatial information from social media feeds". In: *GeoJournal* 78 (2013).

[58] Nicola Stokes et al. "An empirical study of the effects of NLP components on Geographic IR performance". In: *International Journal of Geographical Information Science* 22.3 (2008), pp. 247–264. ISSN: 1365-8816. DOI: 10.1080/13658810701626210.

[59] Benjamin E Teitler et al. “NewsStand: A new view on news”. In: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM. 2008, p. 18.