Semantic Enhancement of Volunteered Geographic Information

Laura Di Rocco

DIBRIS - Università degli Studi di Genova

Abstract. The continue use of social media has created a revolution in the production of data in terms of heterogeneity and volume. This huge amount of data is often georeferenced, and the geospatial position of data is a very relevant dimension for analysis. The considerable heterogeneity and the variable quality of user-generated data prevent however the full exploitation of this information. In this PhD project, we focus on user-generated geospatial data, commonly referred to as Volunteered Geographic Information, where geographical objects are annotated by using tags. The simplicity to use tags allows users to put a lot of tags on objects. This creates a big and noisy tagspace in which it is hard to find data tagged by other people. To overcome this heterogeneity and quality problem, a solution is to rely on ontologies to classify spatial entities tags and names. This problem is addressed in the PhD project with specific reference to the tags used in OpenStreetMap for describing georeferenced objects relevant in urban context, producing as result an ontology for classifying such objects, and a corresponding ontology population approach.

Keywords: Geographic Ontology, Implicit/Explicit Geographical Information, Volunteered Geographical Information, Semantic Classification

1 Introduction

In recent years, we are witnessing a revolution in the production of data, resulting in a significant increase in data complexity, in terms of volume, heterogeneity, and distribution. Data are increasingly being gathered by ubiquitous information-sensing mobile devices, sensor networks, Web applications, software logs. Such a huge amount and variety of data is a valuable source from which to extract information and knowledge. This huge amount of data is often georeferenced, and the geospatial position of data is a very relevant dimension for analysis. The advent of new pervasive communication tools like social media plays a relevant role in such a revolution in data production. Consistent user-generated data represents indeed 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, can be exploited to better understand social dynamics and propagation of information.

Georeferencing information can be explicit, if the user-generated data is explicitly associated with a geographic position –consider for instance the case of content generated from a mobile phone which GPS coordinates are known to the application gathering the content– or implicit, that is it can be deduced, for instance, from the content itself. A typical example is the bulk of geospatial information that can be extracted from short text messages exchanged by users on Twitter. In this case, explicit geographic information can be available in the metadata associated with the tweet (user profile location and GPS coordinates of the device) or 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 users social relationships and activities.

Users play an important role as information producers also for what concerns geospatial information itself, in Volunteered Geographic Information (VGI). Goodchild [5] 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) is by far the most popular. OSM is a collaborative project to create a free editable map of the world via crowdsourced data. The data uploaded to OSM is continuously increasing. However, the considerable heterogeneity and the variable quality of user-generated data prevent the full exploitation of this information. Specifically, in crowdsourced geospatial data, geographical objects are annotated by tags. Tagging seems to be the natural way for people to classify objects as well as an attractive way to discover new material. The simplicity in using tags allows users to put a lot of tags on objects [2]. This results in a big and noisy tagspace in which it is hard to find data tagged by other people. This is due also to the subjectivity of tagging.

To overcome these heterogeneity and quality problems, a solution is to lift from the syntactic to the semantic level both for what concerns spatial entities tags and names, relying on ontologies. The paper is structured as follow: Section 1 presents the problem, Section 2 presents a comparison to state of the art and methodology of our solution and Section 3 presents the case of study and application of our ontology.

2 Problem Statement

The PhD project addresses the problem of classification of explicit geographic information coming from Volunteered Geographic Information. We want to semantically enhance explicit geospatial information available in open source format (OpenStreetMap), thus overcoming the heterogeneity and quality issues inherent in crowdsourced data and tagging discussed above. In order to exploit OSM for georeferencing, OSM tags need to be used.

An alternative approach is proposed by OSMOnto[1] [3]. This is a relevant related project proposing 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. However this work is not relevant for our research as it does not try to correct any possible conceptual mistakes in the taxonomy of OSM tags, but rather to reflect it faithfully in the structure of the ontology.

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

The semantics of OSM tags is quite poor, this work aims at enhancing it through the definition of a properly structured ontology to classify OSM tags. This choice is due to limitation of existing semantic gazetteer. Indeed, classical gazetteer, like Geonames[2], are not very useful as they are too coarse-grained. The proposed approach is thus aimed at extracting accurate and complete georeferenced data (in terms of types of spatial objects) from crowdsourced information, through an appropriate classification. The developed ontology allows to use the data present in OSM as instances of an ontology. We will extract non-spatial information from geospatial data in order to create a classification hierarchy. In this way, OSM users will be able to choose the most suitable tag in OSM for describing a geospatial object. Since the goal of developing a general-purpose ontology for describing geospatial entities is very ambitious and different efforts have been made with different specific goals [7,3,1], we restrict our goal as follows. Our approach aims at classifying geospatial objects that are relevant in urban contexts, thus, that may appear in a generic city, trying to avoid, whenever possible, to focus on the specificities of a particular city. To define our ontology, we keep into account that we mainly want to "semantify" OSM tags, thus we start from the analysis of OSM tags employed in the context of a specific urban context: the city of "London, UK". We start from London (UK) because there is a big amount of OSM data, thus London (UK) provides a good dataset, likely covering most of the relevant concepts in a urban context. Our proposed solution will utilize a classification technique based on "facets". We exploit the possibility to model the domain in three mutual exclusive facets in order to obtain a total generalization.

3 OpenStreetMap Faceted Ontology

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.

State of the art. Several researchers try to extract other kind of information from tweets, for example, to perform sentiment analysis. The main approaches in this field rely on: (a) machine learning techniques, (b) AI techniques, agents and ontologies. For example, the work of Kontopoulos et al. [6] describes two possible approaches to create a domain ontology for sentiment analysis: Formal Concept Analysis and Ontology Learning. In order to create an ontology, we need a collection of objects together with some properties.

For sentiment analysis applications, it is possible to extract this information from a collection of data, i.e., a training dataset.

These two techniques are not suitable for our goal for the following reasons. First of all, because we want to use our ontologies for georeferencing microblogs information (we do not have text to extract information) and secondly because the geographic domain related to a city is not generalizable. Indeed, these two techniques are very good in sentiment analysis domain and in natural language processing field, as we have seen.

[2] http://www.geonames.org/

Methodology. In our work we need to create an ontology that allows us to search explicit geographic information only.

There are multiple ontology engineering methodologies that facilitate the process of developing, maintaining and in overall handling complete life-cycle management. Examples include KACTUS, METHONTOLOGY, SENSUS and DILIGENT. In this work we do not focus on the methodological aspects of ontology development, using rather a simple approach to develop our ontology. This choice is due to the limited complexity of the domain and also because of peculiarity of the project proposal, which requires a manual process to be used.

We do not want to put an over-structure on OSM tags. Ontologies exist that want to put an over-structure on OSM tags. In particular, LinkedGeoData [9] 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.

Our target ontology aims at classifying geospatial objects that are relevant in urban contexts. We cannot exploit existing Gazetteers, like GeoNames, that have a low level of detail respect to internal subdivision of states and are thus too coarse-grained for our purposes. We rather need a semantic gazetteer providing a high level of detail (e.g., inside a city).

We then generalize the ontology classes to encompass concepts that may occur in a generic city, trying to avoid, whenever possible, to focus on the specificities of a particular city. To this aim, we selected the faceted approach to develop the ontology.

Facet ontology [8] classifies 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 facets is designed separately, and models a distinct aspect of the domain. Each facet consists of a terminology, i.e., a finite set of names or terms, structured by a subsumption relation.

In our ontology, facets correspond to geophysical, geopolitical, and Point of Interest aspects. The developed ontology allows data present in OSM to be used as instances of an ontology. A significant semantic support is brought to volunteered geographic data allowing the search at a conceptual level. Conscious of the heterogeneous nature of geospatial data, we do not provide any contribution to the spatial component, already well structured. Instead we aim at improving the non-spatial content which is *per se* heterogeneous and only syntactically semi-structured. The non-spatial content is now accessible through a semantic structure which allows for the conceptual search. The use of "facet" 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.

Population Approach. Differently from what happens for standard ontologies, we decide to manage individuals of our ontology using a different approach. The idea is to have individuals in form of rows of views on a relational database. We will use a new technology able to manage unstructured data with a parallel support. This technology is similar to NoSQL databases but it maintains a relational structure.

The population approach is chosen relying on how our ontology will be used by the future applications. In cases where individuals exist in ontologies, they represent "concrete" objects that have to be classified. However, an ontology need not include any individuals, but one of the general purposes of an ontology is to provide a means of classifying individuals, even if those individuals are not explicitly part of the ontology. For these reasons, we will use database views to manage our individuals. In this way, we are able to directly manage the string type returned by databases. This choice is made to simplify future usage. Indeed, if we have an individual, we have an URI to represent the corresponding object.

With our approach, we perform a matching between OSM objects and our concepts, using simple standard queries on a relational database. With matching we mean the relationship between particular tags and concepts that allow to automatise the population process.

4 Application & Use of Ontology

The developed approach finds a relevant application in the context of our ongoing research project aimed at realtime integration of textual data coming from microblogs and crowdsourced geospatial data [4]. Specifically, the target application first gathers in realtime from Twitter, through the use of the suitable streaming API, both explicitly georeferenced tweets and tweets missing an explicit geospatial information. Then, tweet contents are geoparsed relying on the textual descriptions of OSM objects. Social relationships among users and their activities (such as mentions and retweets) are then exploited to further refine tweet (and corresponding user) geopositioning, since some social relations are strong indicators of spatial proximity. Georeferencing information belonging from different sources (content vs social interaction analysis), appropriately weighted according to the respective confidence, are then merged. Since explicit tagging is used only in a small percentage of tweets, we will use the geospatial information implicit in the messages to improve the resolution of the georeferencing process. This is useful for several applications. For example we could create heat maps to highlight areas from which tweets are generated or areas which tweets refer to. This could be very useful in application such as emergency response.

In our ongoing project [4], we proposed to extract implicit geoinformation contained in tweet contents using a semantic gazetteer. The use of external knowledge like a gazetteer can help us to detect toponyms in the tweet contents. In order to extract toponyms from text, we plan to consider three different approaches:

1. Perform a simple string matching with toponyms extracted from OSM.
2. Improve this 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.

3. 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, ...).

These three approaches can help us to understand and evaluate the extend to which semantic support improves geotagging of non-geotagged tweets.

The semantic enrichment allows us to obtain new geographical knowledge that can be exploited to improve the set of extracted geonames and thus the quality of geotagging.

To evaluate the use of ontology in this type of application, 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 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.

References

1. A. Ballatore and M. Bertolotto. Semantically enriching vgi in support of implicit feedback analysis. In *Web and Wireless Geographical Information Systems*, volume 6574 of *LNCS*, pages 78–93. Springer, 2011.
2. G. Begelman, P. Keller, F. Smadja, et al. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop*, 2006.
3. M. Codescu, G. Horsinka, O. Kutz, T. Mossakowski, and R. Rau. Osmonto-an ontology of openstreetmap tags. *State of the map Europe (SOTM-EU) 2011*, 2011.
4. L. Di Rocco, M. Bertolotto, B. Catania, G. Guerrini, and T. Cosso. Extracting fine-grained implicit georeferencing information from microblogs exploiting crowd-sourced gazetteers and social interactions. In *AGILE International Conference on Geographic Information Science*, 2016.
5. M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
6. E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades. Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074, 2013.
7. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, R. Oltramari, L. Schneider, L. P. Istc-cnr, and I. Horrocks. Wonderweb deliverable d17. the wonderweb library of foundational ontologies and the dolce ontology, 2002.
8. S. R. Ranganathan. Prolegomena to library classification. *The Five Laws of Library Science*, 1967.
9. C. Stadler, J. Lehmann, K. Höffner, and S. Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012.