AN ONTOLOGY-BASED APPROACH TO INCORPORATE USER-GENERATED GEO-CONTENT INTO SDI

Dong-Po Deng$^a$ and Rob Lemmens$^a$

$^a$Department of Geoinformation Processing, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands – {deng, lemmens}@itc.nl

KEY WORDS: Spatial Information Sciences, Technology, Data mining, Query, Understanding, Web based

ABSTRACT:

The Web is changing the way people share and communicate information because of emergence of various Web technologies, which enable people to contribute information on the Web. User-Generated Geo-Content (UGGC) is a potential resource of geographic information. Due to the different production methods, UGGC often cannot fit in geographic information model. There is a semantic gap between UGGC and formal geographic information. To integrate UGGC into geographic information, this study conducts an ontology-based process to bridge this semantic gap. This ontology-based process includes five steps: Collection, Extraction, Formalization, Mapping, and Deployment. In addition, this study implements this process on Twitter messages, which is relevant to the context of professional tasks. It is usually referred to “User-Generated Content (UGC)”. Since UGC often contains explicit or implicit geographic references like GPS coordinates or place names, this content is referred to “User-Generated Geo-Content (UGGC)”. There are often no a priori rules for creating UGGC. The data production of UGGC is a bottom-up approach. In contrast, the data production of Spatial Data Infrastructures (SDIs) is a top-down approach, which is often based on formal geographic knowledge or standards such as feature concept of ISO 19109 General Feature Model.

The different knowledge to create data has resulted in a semantic gap between UGGC and geo-data of SDIs. There is a need to bridge this semantic gap for incorporating UGGC to SDI. Therefore, this study proposes an ontology-based approach to span the semantic gap. By using this ontology-based approach, the ontologies of UGGC can be mapped to formal geographic ontologies of SDI. The ontology mapping can generate integrated ontologies, which can be used to Web applications for understanding people contributions and helping them to create semantically explicit geographic information. This study also implements this ontology-based process for extracting disaster relief information from Twitter’s messages.

Incorporating UGGC into SDIs is becoming an important research task in GI Science. Many researches are considering UGGC as a complementary part of SDIs. Since tremendous people around the world continually contribute UGGC on the Web, people somehow play a role as human sensors to collect data. If UGGC can fit to data model of SDIs, it would reduce the cost to collect geo-data. However, to incorporate UGGC to SDIs is a difficult work. In term of data production, the production of UGGC is different from SDIs’ approach. Both data productions actually are depended on different knowledge.

UGGC is the geo-referenced content of Web. Such content is collaboratively created by people and shared through using Web applications and services. The creation of UGGC is often rooted from practical experiences of people in places, e.g. people who have been a carnival often post a message on the micro-blogging applications like Twitter.

STATE OF THE ART

Ontology is a tool to make knowledge explicit, as well as it is a rudiment for sharing understandings of a particular domain that have to be constructed within social processes among participants (Gruber, 1995). In this context, ontologies can be a method to overcome the problems of semantic heterogeneity in different information resources. Ontology-driven approaches have been applied to minimize the efforts of data integration from distributed and heterogeneous data source (Kolas et al., 2005). To efficiently discover data in such massive and heterogeneous data source, ontologies play an important role.
for modeling metadata schema models and the controlled vocabularies that are used to fill the content of metadata records (Lacasta, et al., 2007). To provide users a semantic interaction with SDI, ontology is used to formally represent conceptual models of the data so that data is consistently visualized, accessed, and processed along the evaluation and exploitation phases (Lutz and Kolas, 2007). To integrate heterogeneous geo-services, ontologies can be used to build a framework of geo-service based on common vocabularies and shared service descriptions (Lemmens et al., 2006).

Ontology mapping is a process to map elements of one ontology to at most one element of another ontology. An ontology mapping can be considered as a collection of mapping rules from one ontology to another. A mapping rule is a correspondence that maps a component of one ontology into another one from another ontology. Correspondence is the relation holding, or supposed to hold according to a particular matching algorithm or individual, between components of different ontologies. These components can be different as classes, individuals, properties or formulas (Euzenat and Shvaiko, 2007).

Ontology mapping is an essential approach for geographic information integration. Kokla and Kavouras (2001) attempt to formalize the geographic context and integrate the existing geographic ontologies into their associated top-level ontologies. The use of method is basically a cross-table to compare different concepts, and then the similar geographic context will be integrated. Kavouras et al. (2005) propose a method to compare the semantic properties and semantic relations of definitions in geographic categories such as land use classifications. The idea of this method is to provide a semantic overview for mapping different geographic ontologies. Stoimenov et al. (2006) develop a hybrid ontology approach based on a semantic mediator architecture for mappings between community terminologies (local ontologies) and mappings between local ontology and top-level ontology.

Generally, an API is set of rules and specifications that an application can access information and use services from another application on the Internet. With having open APIs, many Web 2.0 applications allow users to get information from their databases. OSM API allows users to get the vector data and its attributes. Unlike OSM API, Google Maps API only allows users to overlay their data on raster-based map. Also, UGGC can be received by means of a crawler that uses open APIs provided by Web 2.0 applications. The crawler can either extract multimedia resources together with the tags from Flickr, YouTube, Delicious, etc., or stream textual data from Facebook and Twitter. The multimedia and textual data consisting of implicit geographic knowledge is often inconsistent with professional geographic knowledge. To incorporate UGGC into GI, there is a need to apply an ontology-based process to bridge the semantic gap.

### 3.2 Extraction of structures and patterns from UGGC

This step aims to extract the structures and patterns from UGGC by using data mining or information extraction. Data mining is a common method to extract structures and patterns from low-level data, which are typically too voluminous to understand and digest easily (Fayyad et al, 1996). Information extraction involves the creation of structured representation from filtering information from large volumes of text. Most data mining and information extraction methods are based on statistical methods such as classification rules and clustering algorithms. The development of these structures and patterns is to use these methods to group several similar concepts or link related concepts together. This development is a process to assemble participants’ views to a consensus in an online community. However, the constructions of the patterns and structures are determined by statistical relatedness but not semantic relationships. The lack of logical connections makes it hard to map the elements of the patterns and structures to ontologies. There is a need to improve the patterns and structures to logical, formal representation of knowledge.

### 4.3. Formalization of patterns and structures to community ontologies

This step aims to develop community ontologies from the patterns and structures. In this study, a community ontology is used to describe a lightweight ontology derived by formalizing patterns and structures. In this formalization, online knowledge bases can be used to (1) identity elements of patterns and structures to ontology components such as Class (Concept) or Individual (Instance), and (2) to find the semantic relations between these ontology components such as sub-class relation, part-of, and ad hoc relations of individuals. The online knowledge bases are developed via systematical and logical methods, as well as are available on the Web. For example, Wordnet is an online lexical database of English. It provides short, simple definitions and various semantic relations between the synonym sets. DBpedia is a project that aims to extract structured information from Wikipedia and to make this information accessible on the Web. This structured information covers various domains including geographic information, people, companies, films, music, genes, drugs, books, and scientific publications. The data format of DBpedia is Resource Description Framework (RDF) that allows users can use via SQL-like language (SPARQL) to query the structured information.
3.3 Mapping of community ontologies and application ontologies

The purpose of this step is to map community ontologies to application ontologies. The task of mapping community ontologies to application ontologies uses ontology mapping techniques. An ontology mapping techniques is a process to find a set of correspondences between components of two ontologies. These correspondences can be equivalence relationships, subclass or superclass relationships, transformation rules, and so on (Noy, 2009). While community ontologies are mapped to application ontologies, the application ontologies can identify concepts of UGGC with regarding to geographic knowledge.

Before the ontology mapping, the application ontologies must be developed. As a discussion in Guarino (1998), the application ontologies describe concepts depending on a particular domain knowledge that is often specialized to ontology. The development of the application ontologies has to integrate high-level ontologies such domain ontologies and top-level ontologies. Domain ontologies describe the vocabulary related to a generic domain by specializing the terms introduced in the top-level ontology. Examples of the domain ontology are General Feature Model (GFM) of ISO 19109, Feature Type Thesauri (FTT) of Alexandria Digital Library (ADL) and Ordnance Survey Ontologies. The top-level ontologies describe very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain. For example, Basic Formal Ontology (BFO) (Smith, 2003), Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), and Upper Merged Ontology (SUMO).

Since people create UGGC from any level of concepts, community ontologies would contain abstract and/or concrete concepts. To make the most concepts of communities to be mapped to application ontologies, it is necessary to integrate top-level ontologies and domain ontologies to application ontologies.

4.5. Deployment of application ontologies in Web 2.0 applications

This step focuses on deploying the application ontologies to Web 2.0 applications. This deployment is actually to settle knowledge bases to Web 2.0 applications. The application ontologies can be stored and used in knowledge bases. The development of the knowledge bases is often required the use of semantic Web technologies. To make use of the application ontologies, semantic Web languages such as OWL and RDF are used to semantically represent these ontologies. The use of semantic Web languages makes the knowledge bases possible to query via using SPARQL, as well as reasoning. OWL is designed to support various types of reasoning, typically including subsumption and classification (Hitzler et al., 2009). There are several semantic reasoners supporting OWL such as Pellet (Sirin et al., 2007), RacerPro (Haarslev and Möller, 2003), FaCT++ (Tsarkov and Horrocks, 2006), and Pellet Spatial (Stocker and Sirin, 2009). Moreover, rule languages are developed to fit the various types of reasoning tasks. Semantic Web Rule Language (SWRL) (W3C, 2004) is a rule language that was designed as an extension to OWL. The reasoners and rule-based languages enable the smart query engine to make inference.

4. IMPLEMENTATION OF THE ONTOLOGY-BASED PROCESS ON EXTRACTING GEO-INFORMATION FROM TWEETS

4.1 Tweet collection

In this study, the Japan earthquake tweets are mostly collected via Twapperkeeper because the Twitter search API only returns results from the roughly previous four days. TwapperKeeper is a Web service, which allows users to create and export archives of tweets based on specific keywords or hash tags. We collected tweets from these archives for one month since they were created. The archive #eqjp in TwapperKeeper is created by March 14, 2011. Thus, we collected tweets containing hash tag “#eqjp” from March 11, 2011 to March 14, 2011 via the tweet archive of Digital Library Research Laboratory, Virginia Polytechnic Institute and State University. Figure 2 illustrates a tweet reporting disaster relief information. The data set of Japan earthquake tweets consists of 192,176 tweets and 102,559 twitterers (twitter users). In this data set, there are only 4,234 tweets containing latitude and longitude of a point. The percentage of geo-referenced tweets is quite low. To place a tweet in a location, place names are used to link the tweets to locations. Stanford Named Entity Recognizer (NER) is used to extract locations from the tweets.

4.2 Extraction of triple-like structures from tweets

Tweets are natural-language texts. Twitter users could use any vocabularies to post messages. To understand the vocabularies, keyword co-occurrence analysis is used to explore and examine the vocabularies used in Japan disaster tweets. The selection of the keywords depends on the scenario of the above use case. Since Bruce wants to collect the tweets posting where people need vital resources, the “need” is a keyword for collecting tweets. Take Figure 2 as an example. “Minamisoma” and “need” are co-occurred in this tweets, as well as “food”, “gas”, and “toilet paper” all co-occur with “need” in this tweets. To explore and examine the more vocabularies of disaster relief information in tweets, several keywords are selected such as “need”, “supply”, “open”, “closed”, “damage” and “missing”.

The keyword “need” and “supply” intuitively should co-occur with vital resources. The keyword “open”, “closed”, and “damage” should appear in the tweets containing infrastructures such as roads, schools, and stations. As people post tweets to report missing people, “missing” is a keyword of these tweets.

\[
P(w_x | w_y) = \frac{w_x \cap w_y}{w_y}
\]

Eq. 1 is used to measure co-occurrence relatedness between keywords and vocabularies and determine their hypernym and hyponym relationships. Through the frequency of vocabularies
co-occurring with the keywords in a tweet set, the co-occurrence relatedness between the keywords and vocabularies can develop a concept hierarchy.

In Eq 1, the concept of Word $w_y$ subsumes the concept of Word $w_x$ if $P(w_x | w_y) \geq \alpha$ and $P(w_y | w_x) \leq 1$. The co-occurrence threshold $\alpha$ is empirically determined between 0.7 and 0.8 (Sanderson and Croft, 1999). As the word pairs are selected by using the co-occurrence threshold, a tree graph of super-concept and sub-concept relationships can be developed. But the tree graph often contains redundant relationships. For example, there is a Word $w_x$ that has two potential super-concept Word $w_y$ and Word $w_z$. If Word $w_y$ is also a potential super-concept of Word $w_z$, the super-concept/sub-concept relationship between Word $w_x$ and Word $w_y$ would be removed (Schimitz, 2006).

![Figure 3. A concept hierarchy of keyword “need”](image)

By using the above method, the concept hierarchies of these keywords can be developed. Concept hierarchy is tree structure, which can basically, simply express knowledge. In a concept hierarchy, each node is a concept that is represented by a keyword or a vocabulary, and the relations between super-concept and sub-concept are hyponymy (is-a) and/or meronymy (has-a). The super-concepts and sub-concepts of keywords and vocabularies are determined by the frequency of word co-occurrence in a tweet set. The more tweets a word occurred in, the more general it is assumed to be (Sanderson and Croft, 1999). Figure 3 displays the concept hierarchy of keyword “need”. The vital resources indeed co-occur with the terms “open”, “closed”, “damage”. In addition, place names are extracted from the tweets via Stanford Named Entity Recognizer (NER) in Section 4.1. Thus, the triple-like structures of disaster relief information can be extracted from tweets. Table 1 shows three types of triple-like structures. The Type I is a Resource-Keyword-PlaceName structure. A record of the Type I is extracted if a tweet contains keyword “need” or “supply”, the one of resource vocabularies, and a place name.

<table>
<thead>
<tr>
<th>Type</th>
<th>Vocabularies of Resources/Infrastructures</th>
<th>Keywords</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Food, Water, Clothes, Blanket, Medical, Sanitation, Toilet, paper, Gas, Gasoline, Fuel, Oil, Petrol, Medical, Drug, Vehicle, Volunteer</td>
<td>Need, Supply, Offer, Shortage, Without, No</td>
<td>Place name</td>
</tr>
<tr>
<td>II</td>
<td>Airport, Building, Coast, Expressway, Hospital, Park, Plant, Road, Railway, Station, School, Shelter</td>
<td>Open, Reopen, Closed, Damage</td>
<td>...</td>
</tr>
<tr>
<td>III</td>
<td>Airport, Building, Coast, Expressway, Hospital, Hotel, Park, Plant, Road, Railway, Station, School, Shelter</td>
<td>Open, Reopen, Closed, Damage</td>
<td>Place name</td>
</tr>
</tbody>
</table>

Table 1. The three types of triple-like structures used to extract disaster relief information

4.3 Formalization from triple-like structures to RDF

In this step, the triple-like structure records are transferred to RDF triples. A RDF triple consists of a subject, a predicate, and an object. For the Type I structures, all the vocabularies of the resources are considered as subjects. The keyword “need” and “supply” are transferred to isNeededAt and isSuppliedAt respectively, and their objects are place names. The Type II structures only have two components. The infrastructure names are considered as subjects, and keywords are treated as objects. The predicates between infrastructure names and keywords are hasASubjectStatus. The process of the Type III structure is slightly complicated more than the Type II. A subject of the Type III is a combination of an infrastructure and a place name. Like Type II, the predicates between subjects and keywords are also hasASubjectStatus. To place the Type III structures, it has to add another triple for transfer Type III. The place names in the Type III subjects are used to location identification. The predicates between Type III subjects and place names are isLocatedAt. Figure 4 illustrates the triplified process from the three types of structures to RDF triples.
4.4 Mapping RDF to OGC GeoSPARQL ontology

The GeoSPARQL will be a new OGC standard, which provides three main components for encoding geographic information: (1) The definitions of vocabularies for representing features, geometries, and their relationships; (2) A set of domain-specific, spatial functions for use in SPARQL queries; (3) A set of query transformation rules (OGC, 2011). The ontology of the GeoSPARQL standard includes three main classes: ogc:SpatialObject, ogc:Features, and ogc:Geometry. The ogc:Feature and ogc:Geometry are the subclasses of ogc:SpatialObject. The ogc:Feature class represents features, which are abstractions of real-world phenomena. The concept of feature is derived from ISO 19109 General Feature Model. The ogc:Geometry, expressing spatial geometries of the features, has sixteen subclasses defining a hierarchy of geometry types such as point, polygon, curve, arc, and multicurve. These geometry classes are derived from ISO 19107 Spatial Schema. RDF literals are used to store geometry values. There are two ways to store geometry values via RDF literals: Well Known Text (WKT) and Geography Markup Language (GML). The ogc:asWKT and ogc:asGML properties between the geometry entities and the geometry literals. Geometry values for these two properties use the ogc:WKTLiteral and ogc:GMLLiteral data types respectively.

4.5 Deployment of RDF triples

After mapping RDF triples to the ontology of OGC GeoSPARQL standard, these processed RDF triples have to be stored in a knowledge base for querying. This study uses Parliament, which is an open source triple store developed by Raytheon BBN Technologies (Kolas et al., 2009). The BBN Parliament is compliant with OGC GeoSPARQL standard, and supports spatial and non-spatial SPARQL queries. The spatial queries such as within, buffer, and intersection actually are originated from ISO 19125 Simple Feature Access. By using SPARQL filter functions for expressing spatial queries, Parliament can implement the spatial functions on RDF triples and return results. Figure 7 illustrates how to make a GeoSPARQL query to get the food-needed locations. Figure 8 visualizes the query results on the map.

5. CONCLUSIONS AND FUTURE WORK

To integrate UGGC into SDI, this study proposes an ontology-based process to deal with UGGC. By running this process, geographic information can be extracted from UGGC. As the semantics of UGGC are disambiguated through formal geographic knowledge, the production of geographic
information is not only specialists but also amateurs. This process makes geographic information collection much easy. With using geographic knowledge base, the result of this study can provide users semantically rich queries for geographic information retrieval.

In future work, this ontology-based process will be improved to apply in real time UGGC stream. To achieve this task, the advanced natural language processing techniques to understand the more disaster relief vocabularies in UGGC should be very useful. Moreover, as the more geographic vocabularies from amateurs are identified, it is possible to build an interface, which automatically understands users’ vocabularies. Such interface should be able to make users to contribute geographic information easy.

REFERENCE


