Enriching Top-down Geo-ontologies Using Bottom-up Knowledge Mined from Linked Data

Yingjie Hu and Krzysztof Janowicz

STKO Lab, Department of Geography, University of California Santa Barbara, Santa Barbara, CA 93106, USA

Abstract: Geo-ontologies provide formal specifications of geographic concepts, and can be embedded into geographic information systems to support automatic reasoning. Traditionally, geo-ontologies are developed through a top-down approach in which a group of experts collaboratively decide about the formalization. While such an approach captures valuable expert knowledge, the resulting geo-ontologies could be biased, miss certain useful properties, or may not reflect existing data (needs). The fast evolving Linked Open Data (LOD) cloud offers a large amount of structured data contributed by authoritative agencies, companies, and the general public. With the diverse perspectives and the structured data organization, the LOD cloud contains knowledge which could be used to enrich top-down geo-ontologies. This paper proposes a workflow to mine bottom-up geographic knowledge from the LOD cloud. We describe each step of this workflow, and conduct an experiment using a dataset from the LOD cloud to learn a geographic concept *port city*. We perform an evaluation and show that the workflow can extract useful knowledge for enriching top-down geo-ontologies.

Keywords: geo-ontology, ontology engineering, concept learning, Linked Open Data, Semantic Web, semantics, DBpedia.

1 Introduction

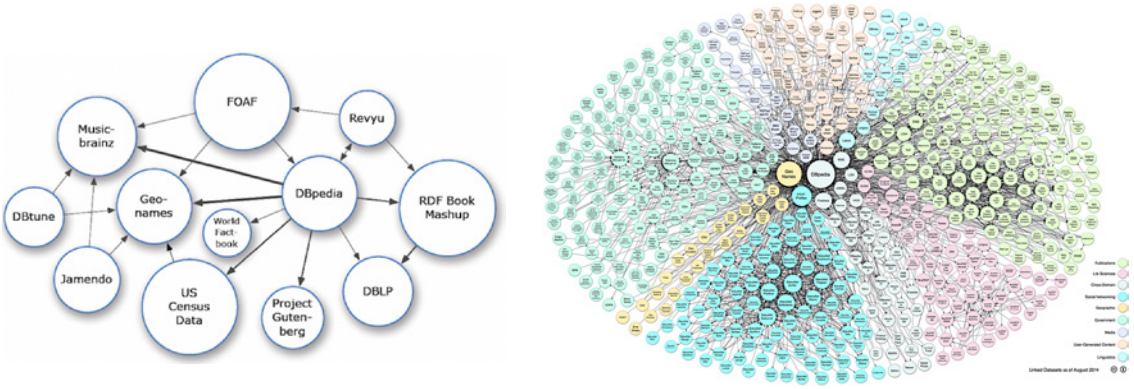
Geo-ontologies provide formal specifications of geographic concepts, and have been discussed in a variety of GIScience studies. As concept mediators, geo-ontologies can enhance the semantic interoperability among heterogeneous data and distributed systems. For example, Fonseca et al. (2002) proposed an architecture which used ontologies as an essential component to integrate

different geographic information (GI) systems [7]. In the domain of environmental monitoring, Pundt and Bishr (2002) developed an ontology to facilitate the sharing of data collected from different field survey activities [20]. Kuhn (2005) proposed *semantic reference systems* which employed ontological specifications to ground and map geographic information in different systems [15]. Geo-ontologies have also been used to improve geographic information retrieval. Jones et al. (2001) combined the semantic relatedness calculated from place ontologies with the traditional Euclidean distance to rank the relevance between the candidate results and the input queries [14]. Li et al. (2011) employed the SWEET ontology to expand the input query with semantically relevant terminologies, thereby enhancing the capability of the traditional keyword-based search [18]. In previous work, we demonstrated how semantic search can be implemented on top of Esri's ArcGIS Online [11]. Capturing expert knowledge, geo-ontologies have also been applied to multiple decision making scenarios. Existing use cases include ontology-driven spatial decision support [17] as well as geodesign [16]. For next generation GI systems, geo-ontologies may play an even more important role by enabling GI systems to automatically recognize geographic entities from data and recommend suitable spatial analysis tools.

Designing good geo-ontologies, however, is not an easy task. Traditionally, a top-down approach has been used, in which a group of experts collaboratively specify the terms and relations of the target ontology. Such an approach has many merits. It captures the valuable domain knowledge from experts, which sometimes can only be acquired after years of experience in the specific field. In addition, the terms assigned by experts are often concise and meaningful since such terms generally have to undergo the deliberations and discussions of multiple professionals. While possessing these merits, the developed geo-ontologies may nevertheless be biased towards the knowledge of the participating experts, and may miss some properties which could be useful for understanding the specific geographic concept, and many not well reflect particular datasets or future use cases.

Progress in Semantic Web technologies [6] fostered the fast evolution of the Linked Open Data (LOD) cloud [2]. From 2007 to 2014, the LOD cloud has grown from its initial 12 datasets to more than 570 datasets with billions of triples (see Figure1). These rich amount of data are contributed by authoritative agencies, such as the U.S. Census and data.gov.uk, the industry, and also the general public. Examples of such user-contributed datasets include DBpedia and LinkedGeoData, which are the Linked Data versions of Wikipedia and OpenStreetMap respectively [1, 25]. Data instances in the LOD cloud are structured using the Resource Description Framework (RDF). This structured organization distinguishes LOD datasets from other unstructured user-contributed content such as most social media data.

The LOD cloud presents a valuable resource from which bottom-up knowledge could be mined to enrich the top-down geo-ontologies. The value of the Linked Data cloud can be seen in two ways. First, with many different data



(a) The LOD cloud in 2007. (b) The LOD cloud in 2014.

Figure 1: Evolution of the Linked Open Data cloud from 2007 to 2014 [22].

contributors, datasets on the LOD cloud reflect the diverse perspectives of people towards the same concepts and entities, and therefore can be exploited to enrich the knowledge from the limited number of participating experts. Second, the structured data enable knowledge to be extracted in a structured manner (e.g., in the form of properties and property-values) which is often desired for an already formalized top-down geo-ontology. However, mining knowledge from the LOD cloud demands suitable methods, since improper approaches (e.g., using a natural language processing method based on a *bag of words* model) may simply break the links among data instances and convert the structured data back to an unstructured form.

This paper is an effort towards extracting bottom-up knowledge from the LOD cloud. The contributions of this work are as follows:

- We develop a workflow that mines knowledge about geographic concepts from the structured Linked Open Data.
- We demonstrate the use of the workflow by applying it to a sample dataset from DBpedia and an example top-down geo-ontology.
- We designed a preliminary experiment to evaluate the extracted bottom-up geographic knowledge.

The remainder of this paper is organized as follows. Section 2 reviews related work on geo-ontology engineering, and provides some background on Linked Data and DBpedia. Section 3 presents the workflow for extracting geographic knowledge from Linked Open Data. In section 4, we employ the proposed workflow to mine knowledge from DBpedia, and perform a preliminary evaluation on the extracted knowledge. Finally, section 5 summarizes our work and discusses future directions.

2 Related work

The value of geo-ontologies has long been recognized by the GIScience community, and the history can be traced back to a NCGIA specialist meeting in 1998 [19]. Unlike the ontology discussed in philosophy, geo-ontologies are closer to those in computer science, which are designed to help machines turn data into sharable knowledge [4, 9]. Different from ontologies in other domains (e.g., bioinformatics), geo-ontologies focus on achieving better understanding of the geographic world and facilitating the implementation of conceptually sound GI systems [23]. Since the 1998 meeting, a lot of studies have been devoted to developing geo-ontologies. Smith and Mark (2001) investigated the conceptualization of non-expert subjects on geospatial phenomenon, and derived an ontology of geographical categories [24]. Frank (2003) designed a 5-tier ontology for spatio-temporal databases which starts from the observations in the physical world and completes at the knowledge of cognitive agents [8]. Scheider et al. (2009) developed a formalization for grounding geo-ontologies in the physical environment [21]. Focusing on geographic information constructs, Couclelis (2010) developed a hierarchical framework with the user intentionality on one end and the existence of information on the other [5]. Janowicz (2012) proposed an observation-driven ontology engineering framework which aims at deriving ontological primitives from observation data [12]. The work at hand has been influenced by these previous studies. However, we focus on extracting bottom-up geographic knowledge from Linked Data to enrich top-down geo-ontologies, which has been rarely examined so far.

The growth of the LOD cloud brings a large amount of structured spatiotemporal data, and is changing the ways of publishing, searching, and sharing geographic information [13]. The term *Linked Data* has two folds of meanings that are often used interchangeably. On the one hand, it refers to a set of principles recommended by W3C for publishing data on the Semantic Web. On the other hand, it represents the data which are structured and published following these principles. Among the many datasets on the LOD cloud, DBpedia is a central hub, which provides information about more than 4.5 million entities (many of which are geographic places) [3]. The content of DBpedia originates from Wikipedia, and each Wikipedia article has a corresponding DBpedia page. As a result, DBpedia inherits many great features of Wikipedia. For example, Wikipedia articles are contributed by over 25, 272, 000 users (http://en.wikipedia.org/wiki/Wikipedia:Statistics, retrieved in May 2015), and accordingly, DBpedia data obtain the diverse perspectives from the huge number of people. Meanwhile, a lot of data on Wikipedia have their original sources from authoritative agencies. For example, by examining the Wikipedia page of *San Francisco*, one can find that the data about the city's land and water areas come from U.S. Census, while the elevation data are from U.S. Geological Survey. Unsurprisingly, DBpedia also inherits these valuable authoritative data. Since new content is being constantly added to Wikipedia, DBpedia updates its data regularly to synchronize with Wikipedia.

Categorization systems are frequently used by datasets on the LOD cloud to group similar instances. In contrast to LOD datasets that employ pre-defined categorization schemata, Wikipedia allows voluntary contributors to create customized categories and to classify entities into these categories. For example, there exists a category called *Port cities and towns of the United States Pacific coast* (see http://en.wikipedia.org/wiki/Category:Port_cities_and_towns_of_the_United_States_Pacific_coast) which contains cities, such as *San Francisco* and *Los Angeles*. According to Wikipedia, the intention of these categories is to "*group together pages on similar subjects*". To some degree, the categorization result is similar to the data that Smith and Mark (2001) collected in [24]. In their experiment, non-expert human subjects were asked to give examples for geographic categories, whereas Wikipedia invites users to perform categorization tasks on the Web. DBpedia inherits these customized categories and the classification results from Wikipedia. In this work, we make use of the data instances under specific geographic categories to discover the properties which differentiate the instances that are in a category from those that are not.

3 Workflow

The objective of the proposed workflow is to extract geographic knowledge from Linked Open Data in a bottom-up manner. Specifically, we aim at discovering the knowledge which may be missing or biased in top-down geo-ontologies. The top- down geo-ontologies discussed in this paper are not the top-level ontologies in existing literature, which provide abstract and domain-independent terms, such as *endurant* and *perdurant*. Instead, these top-down geo-ontologies model concrete geographic concepts (e.g., *lake* and *university town*), and are micro-ontologies which serve as building blocks in specific applications [13]. Figure 2 shows an overview of the designed workflow.

The workflow starts from the *Linked Datasets* at the lower-left corner of the figure. Based on the categorization system, the workflow first selects a *Target category* that corresponds to the geographic concept modelled by the top-down geo-ontology. Meanwhile, both *positive instances* and *negative instances* are selected according to the category. Positive instances are the entities which are classified by users as belonging to the target category, whereas negative instances are those that do not belong to the category. For example, if the target category is *university town*, then positive instances are the towns which have been classified into this category, while negative instances are those which are not considered as *university town*. Selecting suitable positive and negative instances are important since they will be used as the input for the next three-stage process to learn the target category.

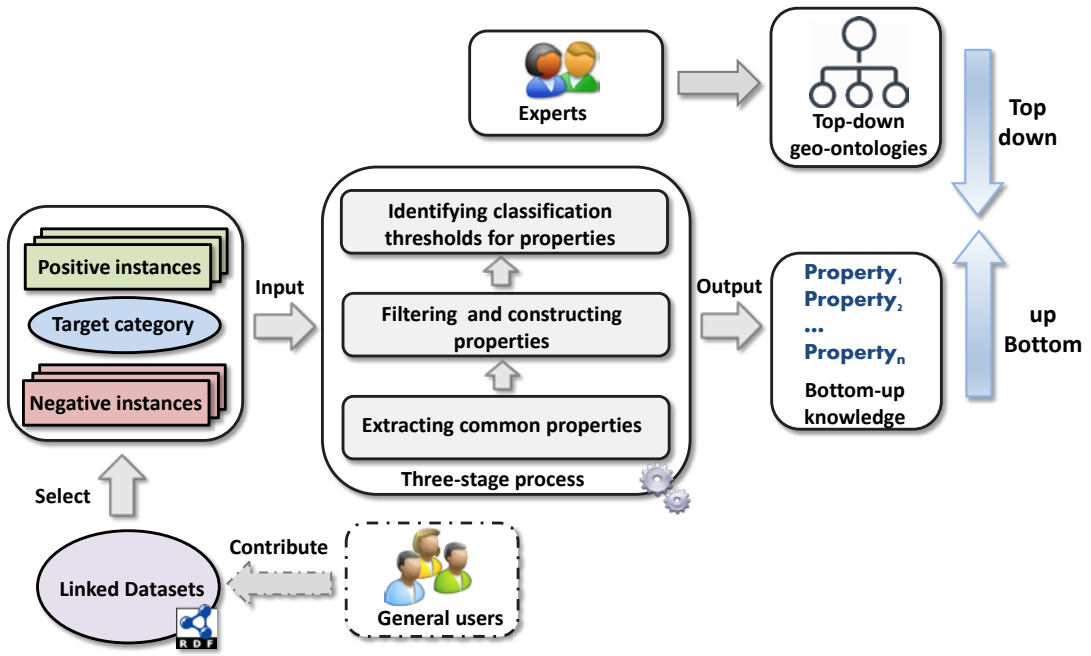


Figure 2: An overview of the workflow. The dotted lines of the *General users* box and the *Contribute* arrow represent that some Linked Datasets are not contributed by general users).

The first stage extracts common properties among the input instances. It examines all the properties that each instance has, and attempts to answer two questions: 1) what are the property-value pairs that only exist in the *majority* of the positive instances? 2) what are the properties that are shared by both the positive and negative instances? Answering the first question can help discover the properties whose existence indicates a strong membership between a geographic instance and the target category. For example, an examination of the instances in the category of *university town* may reveal that the property-value pair (i.e., a *predicate* and *object* pair) *hasUniversity.University* is shared among the majority of the positive instances while not in most of the negative instances. Such a result indicates that this property-value pair is a strong indication for an instance to be considered as a *university town*. The term *majority* should be determined based on the requirements of specific applications. For example, if the goal is to learn a category that is compatible to a few outliers, then a value of 95% could be used as the *majority* threshold, and it means the properties are shared by at least 95% of the positive instances and by no more than 5% of the negative instances. Answering the second question can help find the candidate properties whose value ranges can be potentially used to distinguish a geographic concept. For example, to learn the concept *big city* in the mind of the general public, a property *population* may be shared by both positive and negative instances. While this property is not unique to positive instances, its value can still be used to differentiate the target category (e.g., a *big city* might have *population* $> 1,000,000$ based on the user-contributed data).

The second stage filters out certain irrelevant properties and constructs new properties which might be useful for understanding the target concept. This stage is a supervised process and requires manual intervention. The reason that this filtering is necessary is because the LOD cloud is a big knowledge base that is not merely for one specific application: the information available on the LOD cloud is much richer than what is typically needed for an application. Thus, instead of directly picking and using the data, applications should be selective in terms of what data are relevant and what are not. When it comes to learning knowledge about a specific geographic concept, some properties may not be considered as relevant. For example, in DBpedia, a geographic place is often linked to the celebrities who were born there through the property *isHometownOf*. Such a property can be useful in understanding the relations between people and places, and in fact, we have utilized these relations from DBpedia to improve place name disambiguation in a previous study [10]. However, this property may not be relevant if the concept we want to learn is *university town*. One may wonder why this property filtering is not put into the first stage to pre-process the data. This is due to the manual work it requires: removing the irrelevant properties after the common ones have been identified can save a lot of human effort. In addition, new properties can be constructed based on the existing ones. For example, if both *total area* and *total population* about a city are available in the data, one can construct a new property *population density*, which may become very valuable information in identifying some geographic concepts, such as *populous city*.

The third stage examines the properties that are the output from the second stage. This stage also answers two questions for each examined property: 1) whether this property can be used to differentiate positive and negative instances? 2) if yes, what is a suitable threshold for this property to separate data instances? Before the more detailed method is presented, let us first consider two example properties (see Figure 3). Intuitively, the property in Figure 3(a) can be used to differentiate positive and negative instances, whereas the property in Figure 3(b) cannot since its instances are mixed together.

(a) A property with a clear cut. (b) A property with mixed instances.

Figure 3: Two example properties. Green circles are positive instances and red circles are negative ones. The horizontal arrow indicates the increasing direction of the property values.

In order to find the properties similar to Figure 3(a), we use a method based on entropy and information gain. Entropy is a metric which quantifies the randomness of information [6], which can be calculated using equation 3.1.

$$entropy(X) = -\sum_{i=\{pos,neg\}} P(x_i) \log P(x_i) \quad (3.1)$$

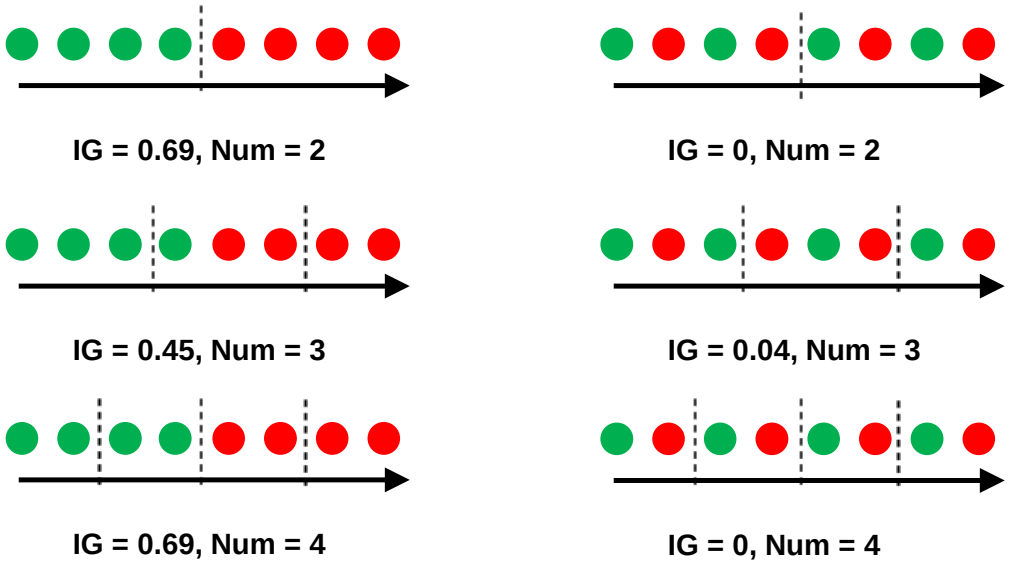
where $entropy(X)$ represents the entropy of the dataset X, x_{pos} represents the positive instances in X, and x_{neg} represents the negative instances. $P(x_i)$ is the empirical proportion of either positive or negative instances in the dataset, which can be calculated, for example, by $x_{pos}/(x_{pos} + x_{neg})$ for positive instances.

Information gain (IG) is the entropy difference before and after an action has been performed on the data. It can be calculated using equation 3.2.

$$IG = entropy_b(X) - entropy_a(X) \quad (3.2)$$

where IG represents the information gain, $entropy_b(X)$ is the entropy before applying the action (which is regular segmentation in this work), and $entropy_a(X)$ is the entropy after the action.

Our method integrates entropy, IG, and regular segmentation to examine the properties output from the second stage. For each property, we segment the data instances evenly into multiple groups, and calculate the information gain by subtracting the entropies before and after the segmentation. We perform this process iteratively with an increasing number of segmentations. The rationale behind this method is that for the properties that have a clear cut, their information gain will increase quickly and will soon reach a plateau with the increasing number of segmentations, since most of the segmented groups will contain only one type of instances; on the contrary, for the properties that have mixed instances, their information gain will not show such a rapid increase, since larger number of segmentations still cannot separate the positive and negative instances. Figure 4 illustrates this process by applying an increasing number of segmentations to the two example properties shown in Figure 3.



(a) A property with a clear cut. (b) A property with mixed instances.

Figure 4: Information gains with different numbers of segmentation.

It can be seen that the property in Figure 4(a) quickly reaches its maximum IG value when the number of segmentation is 2. Although there are fluctuations, this property still achieves very high IG. In contrast, the IG value of the property in Figure 4(b) only increases slowly with fluctuations. Both properties will reach their highest IG values, when the number of segmentations becomes extremely high, in which each separated group contains only one single instance. By plotting out the relation between IG and the number of segmentations, we can visually identify those properties which quickly reach their plateaus. Examples of such plots will be shown in the following section 4. After these properties have been identified, their suitable value ranges can be extracted by aggregating the values of the major positive instances. Similarly, a *majority* threshold, such as 95%, could be used to make the learned concept compatible to a few outliers.

4 Experiment

This section describes an initial experiment which uses the proposed workflow to learn the geographic concept *port city* from the DBpedia data.

4.1 Experimental data and geo-ontology

A top-down geo-ontology constructed to model this concept can be in the form of Figure 5. In this ontology, a *port city* inherits from a super and more general class *city*, and it has a *port* and is close to a *water body*. These are some intuitive properties that make a *city* as a *port city*.

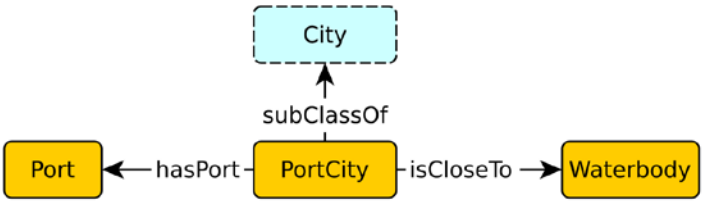


Figure 5: A simplified example top-down geo-ontology for *port city*. The light blue rectangle represents the super class defined in an existing geo-ontology, and the yellow rectangles represent the classes defined in this ontology.

To learn bottom-up geographic knowledge about this concept, we can follow the presented workflow. First, a *target category* needs to be identified. In this experiment, two categories from DBpedia have been used, which are *Port cities and towns of the United States Atlantic coast* and *Port cities and towns of the United States Pacific coast*. The cities belonging to these two categories are combined into one set as the positive instances. In total, 49 positive cities have been identified, and all of their properties have been retrieved from DBpedia. It is worth noting that these 49 cities do not cover all port cities in the U.S., and some cities, such as *New Orleans*, can be well considered as port cities. However, these 49 cities have been

explicitly classified by Wikipedia users as port cities, and therefore have been used as the training data. For negative instances, since DBpedia does not provide the data on which cities are not port cities, 29 inland U.S. cities have been randomly selected. Figure 6 shows the geographic distribution of the selected cities. As can be seen, the port cities used in this experiment are distributed along the east and west coasts, while the non-port cities are located inside the U.S. continent.

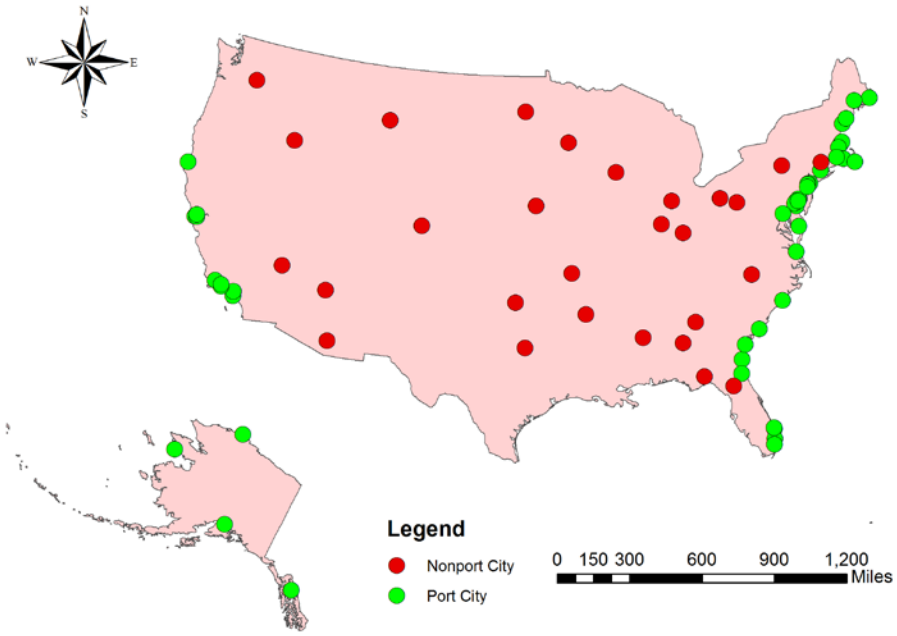


Figure 6: Geographic distribution of the port and non-port cities. Green circles are positive instances and red circles are negative instances.

4.2 Experimental procedure

Extracting common properties. A 95% threshold has been used to extract the common properties. The extraction process takes two steps. First, we examine all data to identify the properties that are shared by both positive and negative instances. The identified properties are shown in Table 1. As can be seen, many properties provide useful information about the cities, such as their populations, land areas, related roads, companies, and other information. In the second step, the same 95% threshold has been applied to only positive instances to extract the distinctive properties. In addition to the properties shown in Table 1, one more property-value pair was extracted, which is *is dbpedia-owl:homeport of* whose value is *dbpedia:Ship*. This result is consistent with what has been defined in the top-down geo-ontology: a *PortCity* should have a *Port*, and accordingly should be the *homeport* of something, such as *Ship*. This consistence demonstrates that the properties developed in a top-down approach can be confirmed by the bottom-up knowledge extracted from the data.

dbpedia-owl:areaTotal	is dbpedia-owl:city of
is dbpedia-owl:premierePlace of	dbpedia-owl:leaderName
is dbpedia-owl:routeStart of	dbpedia-owl:foundingDate
is dbpedia-owl:location of	dbpedia-owl:leaderTitle
dbpedia-owl:areaLand	is dbpedia-owl:builder of
dbpedia-owl:isPartOf	is dbpedia-owl:locationCity of
is dbpedia-owl:assembly of	dbpedia-owl:type
is dbpedia-owl:headquarter of	is dbpedia-owl:foundationPlace of
dbpedia-owl:routeEnd	is dbpedia-owl:region of
dbpedia-owl:owner	is dbpedia-owl:residence of
is dbpedia-owl:nearestCity of	is dbpedia-owl:broadcastArea of
dbpedia-owl:abstract	is dbpedia-owl:populationPlace of
dbpedia-owl:populationTotal	is dbpedia-owl:broadcastArea of
is dbpedia-owl:deathPlace of	dbpedia-owl:timeZone
is dbpedia-owl:routeJunction of	dbpedia-owl:locatedInArea
dbpedia-owl:utcOffset	is dbpedia-owl:hometown of
dbpedia-owl:areaWater	is dbpedia-owl:restingPlace of
is dbpedia-owl:birthPlace of	dbpedia-owl:elevation
geo:lng geo:lat	dbpedia-owl:postalCode
dbpedia-owl:areaCode	

Table 1: Properties shared by 95% of both the positive and negative instances.

Filtering out irrelevant properties and constructing new properties. This stage removes the irrelevant properties and constructs new ones for learning the concept *port city*. The irrelevant properties have been classified into the following categories:

- Linking to persons related to the place, e.g., *is residence of*, *is deathPlace of*, *is home- Town of*, *is restingPlace of*, *is birthPlace of*, ...
- Linking to organizations at the place, e.g., *is city of*, *is location of*, *is headquarter of*, *is foundationPlace of*, *is broadcastArea of*, ...
- Linking to roads and highways, e.g., *is routeStart of*, *is routeEnd of*, ...
- Linking to political or administrative information, e.g., *leaderName*, *leaderTitle*, *postCode*, *areaCode*, ...

After filtering out these irrelevant properties, the rest are summarized in Table 2. Although only 5 properties remain, they all convey important geographic information about the places. In addition to the 5 properties, one new property, *waterLandPercentage*, has been constructed which is calculated by *areaWater/areaTotal*. This new property is added since it can be directly relevant to the concept of *port city*. These properties will be tested in the next stage to see if they can be used to differentiate the positive and negative instances.

dbpedia-owl:areaTotal	dbpedia-owl:areaLand	dbpedia-owl:areaWater
dbpedia-owl:populationTotal	dbpedia-owl:elevation	waterLandPercentage

Table 2: Properties output from the second stage.

Identifying classification thresholds for properties. Regular segmentation, entropy, and information gain have been applied to the 6 properties in Table 2. Figure 7 shows the plotted results with the number of segmentations on the x axis, and the values of IG on the y axis. As can be seen, with the increase of the segmentation number, the information gains of different properties increase in different manners. For some properties, such as the *elevation* (Figure 7(e)) and the *waterLandPercentage* (Figure 7(f))), their information gains increase rapidly and reach the plateau soon. These are the properties which can effectively separate positive and negative instances. The other 4 properties, in contrast, show slow increases and constant fluctuations with different segmentations. This result indicates that these properties have mixed positive and negative instances, and therefore are not suitable for learning the concept *port city*.

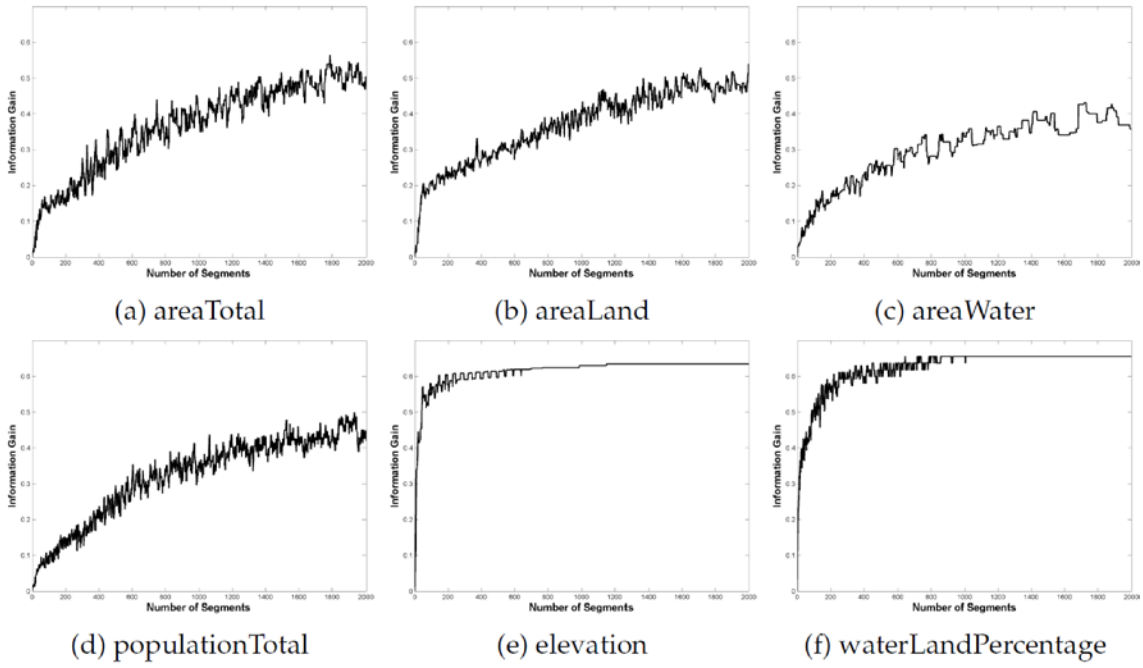


Figure 7: Plots of information gain and segmentation numbers for different properties.

For the two identified properties, *elevation* and *waterLandPercentage*, the values of 95% of the positive instances are aggregated, and the obtained threshold results are listed as below (the unit of elevation is *meter*).

$$elevation < 49.36 \tag{4.1}$$

$$waterLandPercentage > 11.79\% \tag{4.2}$$

The extracted knowledge about *port city* is reasonable: generally, a port city is located at places where the average elevation is not too high and which have quite an amount of water within the city boundary. However, such knowledge could be missed during a top-down ontology development process.

4.3 An evaluation of the extracted knowledge

To evaluate the quality of the learned geographic knowledge, an unseen DBpedia dataset has been used. This dataset contains 21 German cities which have been classified by Wikipedia users into the category *port cities* in *Germany*, as well as 17 cities randomly selected from the inland of Germany as the negative instances. The geographic distribution of the positive and negative instances in this testing dataset is shown in Figure 8.

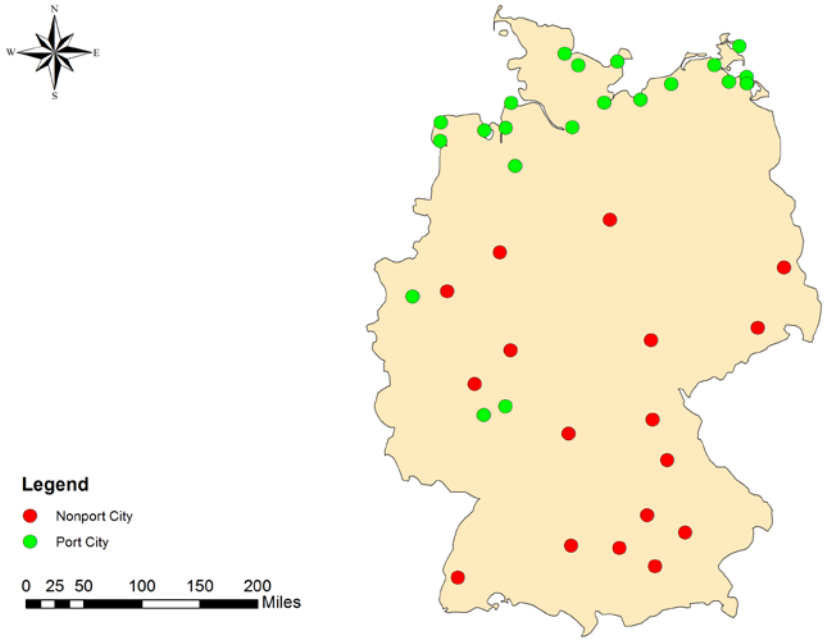


Figure 8: Geographic distribution of the positive and negative instances in the testing dataset (green circles are positive instances and red circles are negative ones).

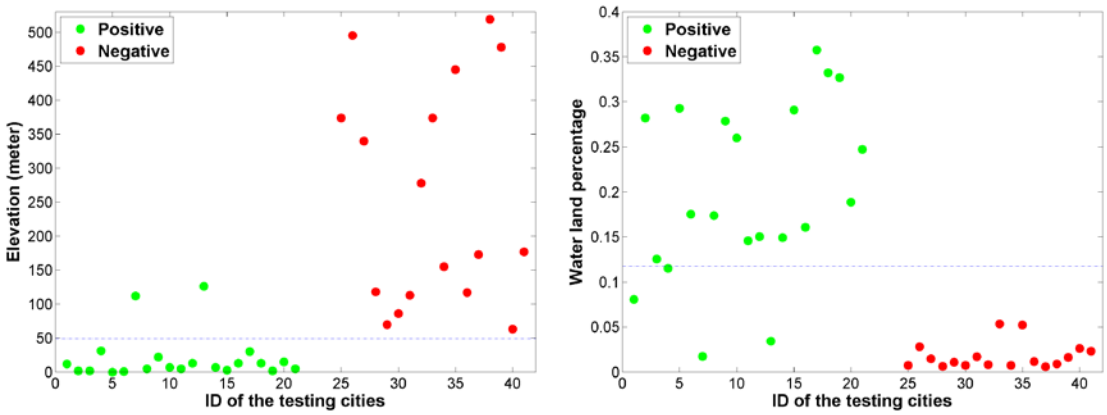
It can be seen that two of these positive port cities, namely *Frankfurt* and *Mainz*, lie in the inland of the country. These two cities are along the *Main river*, and have been generally considered as *river port* cities (in contrast to the other *seaport* cities). This difference can help ontology developers rethink and refine the target concept to meet the application requirement.

The two pieces of mined knowledge in equations 4.1 and 4.2 are examined using the metric of *accuracy* from information retrieval, which is defined in equation 4.3.

$$accuracy = (TP + TN)/(P + N) \tag{4.3}$$

where *TP* represents *true positive* which are the number of positive instances that are also considered as positive by the extracted knowledge. For example, if a port city (positive instance) has an average *elevation* lower than 49.36 meters as

learned from our experiment, then this instance will be counted into *TP*. Similarly, *TN* represents *true negative* which are the number of negative instances that are also considered as negative by the extracted knowledge. For example, if a non-port city (negative instance) has a *waterLandPerventage* lower than 11.79% (thus, it is correctly considered as a non-port city), then this instance will be counted into *TN*. *P* and *N* are the total numbers of positive and negative instances in the testing dataset. The metric *accuracy* measures the consistency between the extracted knowledge and the unseen testing instances.



(a) elevation (b) waterLandPercentage

Figure 9: Evaluation of the testing cities in Germany based on the two extracted properties. Green circles are positive instances, and red circles are negative ones. The dotted line represents the reference value based on the extracted knowledge.

The knowledge about *elevation* is first evaluated against the testing data (see Figure 9(a)), and the following result is acquired: $TP/P: 19/21, TN/N: 17/17$, *accuracy*: 94.74%. It can be seen that the geographic knowledge learned about *elevation* is highly consistent with the testing data. The two cities which are classified incorrectly are the two inland port cities.

When it comes to evaluating the knowledge on *waterLandPercentage*, there exists a challenge: the *Germany* cities in the testing DBpedia dataset do not have the property of *areaWater* which is necessary in this experiment to calculate the *waterLandPercentage*. Such a situation can be attributed to the varied data availability in different countries. In order to test this extracted knowledge, we make use of the geographic data from OpenStreetMap. The administrative boundaries and the water-related areas (such as *river*, *lake*, *bay*, and *reservoir*) have been downloaded for each of the testing cities. We sum up the water areas and the administrative areas respectively, and then calculate the water land percentages by dividing the former with the latter. The calculated values are plotted out in Figure 9(b). By applying the knowledge *waterLandPercentage* > 11.79% to the testing cities, we obtain the following result: $TP/P: 17/21, TN/N: 17/17$, *accuracy* : 89.47%.

5 Conclusions and future work

Geo-ontologies can play an even more important role in developing the next generation intelligent GIS by enabling the systems to automatically recognize geographic concepts from data and recommend suitable tools. While a top-down approach has often been used to develop geo-ontologies, such an approach may be biased towards the knowledge of the participants or miss some useful properties. The fast growth of the Linked Open Data cloud provides a valuable resource for deriving knowledge in a bottom-up manner. Such knowledge can then be used to enrich and complement the top-down geo-ontologies. This paper presents early results about a workflow for mining bottom-up geographic knowledge from Linked Open Data. Based on both positive and negative instances of a target concept, the workflow identifies the common properties, filters irrelevant information, and extracts suitable thresholds. An initial experiment has been conducted, in which the workflow has been used to extract knowledge about a geographic concept *port city* from a DBpedia dataset. We evaluate the extracted knowledge using an unseen dataset, and the evaluation result shows a good consistency between the learned knowledge and the testing cities. While DBpedia has been used as the data source in the experiment, the proposed workflow can also be applied to other LOD datasets.

The performance of the proposed workflow depends on the availability and quality of the training data which contain the target category, the positive instances, and the negative ones. While we obtained our data from the Wikipedia categorization system in this work, other approaches could also be used. For example, traditional human participant experiments could be employed to elicit the typical instances of a target category. The derived instance memberships can then be embedded into the presented workflow, and combined with the LOD datasets to mine bottom-up knowledge. Alternatively, one can create the target category on Wikipedia, encourage online users to classify instances based on this category, and then harvest the data. While the latter approach might require less human effort and thus better scale up, traditional human participant tests provide more information about the background of the participants (e.g., age and gender), and therefore can provide a more representative data sample.

The presented research can also be extended in several directions. First, our experiment so far has examined the applicability of the proposed workflow using one geographic concept. While fair performance has been observed, it is still necessary to investigate some additional concepts to understand the merits and limitations of the proposed workflow more thoroughly. Such investigation could also help quantify the degree of improvement that our workflow can bring to existing top-down geo- ontologies. Second, the evaluation experiment indicates that the *port cities* in the U.S. are similar to the *port cities* in Germany in terms of their elevations and water land percentages. This

result is intriguing since some geographic concepts (e.g., *mountain* and *hill*) may be conceptualized differently in different countries. Thus, it would also be interesting to examine which concepts are more regionally sensitive and which others are more stable across different geographic areas.

References

[1] AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., AND IVES, Z. DBpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 2007, pp. 722–735.

[2] BIZER, C., HEATH, T., AND BERNERS-LEE, T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems 5*, 3 (2009), 1–22.

[3] BIZER, C., LEHMANN, J., KOBILAROV, G., AUER, S., BECKER, C., CYGANIAK, R., AND HELLMANN, S. Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web 7*, 3 (2009), 154–165.

[4] COUCLELIS, H. Ontology, epistemology, teleology: triangulating geographic in- formation science. In *Research trends in geographic information science*. Springer, 2009, pp. 3–15.

[5] COUCLELIS, H. Ontologies of geographic information. *International Journal of Geographical Information Science 24*, 12 (2010), 1785–1809.

[6] FLACH, P. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[7] FONSECA, F. T., EGENHOFER, M. J., AGOURIS, P., AND CÂMARA, G. Using ontologies for integrated geographic information systems. *Transactions in GIS 6*, 3 (2002), 231–257.

[8] FRANK, A. U. *Ontology for spatio-temporal databases*. Springer, 2003.

[9] GUARINO, N. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, vol. 46. IOS press, 1998.

[10] HU, Y., JANOWICZ, K., AND PRASAD, S. Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. In *Proceedings of the 8th workshop on geographic information retrieval* (2014), ACM New York, NY, pp. 1–8.

[11] HU, Y., JANOWICZ, K., PRASAD, S., AND GAO, S. Enabling semantic search and knowledge discovery for arcgis online: A linked-data-driven approach. In *AGILE 2015*. Springer, 2015, pp. 107–124.

[12] JANOWICZ, K. Observation-driven geo-ontology engineering. *Transactions in GIS 16*, 3 (2012), 351–374.

[13] JANOWICZ, K., SCHEIDER, S., PEHLE, T., AND HART, G. Geospatial semantics and linked spatiotemporal data–past, present, and future. *Semantic Web 3*, 4 (2012), 321–332.

[14] JONES, C. B., ALANI, H., AND TUDHOPE, D. Geographical information retrieval with ontologies of place. In *Spatial information theory* (2001), Springer, pp. 322–335.

[15] KUHN, W. Geospatial semantics: why, of what, and how? In *Journal on Data Semantics III*. Springer, 2005, pp. 1–24.

[16] LI, N., ERVIN, S., FLAXMAN, M., GOODCHILD, M., AND STEINITZ, C. Design and application of an ontology for geodesign. *Revue internationale de géomatique*, 22 (2012), 145–168.

[17] LI, N., RASKIN, R., GOODCHILD, M., AND JANOWICZ, K. An ontology-driven framework and web portal for spatial decision support. *Transactions in GIS 16*, 3 (2012), 313–329.

[18] LI, W., YANG, C., NEBERT, D., RASKIN, R., HOUSER, P., WU, H., AND LI, Z. Semantic-based web service discovery and chaining for building an arctic spatial data infrastructure. *Computers & Geosciences 37*, 11 (2011), 1752–1762.

[19] PEUQUET, D., SMITH, B., AND BROGAARD, B. O. The ontology of fields. Tech. rep., National Center for Geographic Information and Analysis, 1998.

[20] PUNDT, H., AND BISHR, Y. Domain ontologies for data sharing–an example from environmental monitoring using field GIS. *Computers & Geosciences 28*, 1 (2002), 95–102.

[21] SCHEIDER, S., JANOWICZ, K., AND KUHN, W. Grounding geographic categories in the meaningful environment. In *Spatial Information Theory*. Springer, 2009, pp. 69–87.

[22] SCHMACHTENBERG, M., BIZER, C., JENTZSCH, A., AND CYGANIAK, R. Linking open data cloud diagram, http://lod-cloud.net/, 2014.

[23] SMITH, B., AND MARK, D. M. Ontology and geographic kinds. In *Proceedings of the Tenth International Symposium on Spatial Data Handling*, T. Poiker and N. Chris- man, Eds. Simon Fraser University, 1998, pp. 308–320.

[24] SMITH, B., AND MARK, D. M. Geographical categories: an ontological investigation. *International journal of geographical information science 15*, 7 (2001), 591–612.

[25] STADLER, C., LEHMANN, J., HÖFFNER, K., AND AUER, S. LinkedGeoData: A core for a web of spatial open data. *Semantic Web 3*, 4 (2012), 333–354.