A Data-Driven Approach for Detecting and Quantifying Modeling Biases in Geo-Ontologies Using a Discrepancy Index

Bo Yan, Krzysztof Janowicz, and Yingjie Hu

STKO Lab, Department of Geography, University of California, Santa Barbara, USA
{boyan,jano,yingjiehu}@geog.ucsb.edu

Abstract

Geo-ontologies play an important role in fostering the publication, retrieval, reuse, and integration of geographic data within and across domains. The status quo of geo-ontology engineering often follows a centralized top-down approach, namely a group of domain experts collaboratively formalizing key concepts and their relationships. On the one hand, such an approach makes use of the invaluable knowledge and experience of subject matter experts and captures their perception of the world. On the other hand, however, it can introduce biases and ontological commitments that do not well correspond to the *data* that will be semantically lifted using these ontologies. In this work, we propose a data-driven method to calculate a *Discrepancy Index* in order to identify and quantify the potential modeling biases in current geo-ontologies. In other words, instead of trying to measure quality, we determine how much the ontology differs from what would be expected when looking at the data alone.

Keywords: geo-ontology; ontology engineering; DBpedia; Linked Data; Discrepancy Index

1 Introduction

Due to the diverse and eclectic nature of geographic information, geographic data usually comes from different sources, in different formats, and are conceptualized from different perspectives. These heterogeneities in terms of provenance and standards create a barrier for integrating data to perform more comprehensive analysis. Geo-ontologies provide a promising way to alleviate this long-standing issue by enabling a flexible integration of geographic information based on semantics, i.e., regardless of representational choices and syntax.

However, the common ways in which geo-ontologies are developed top-down by a team of knowledge engineers and domain experts carry the risk of generating biased or unsuitable geo-ontologies (Hu and Janowicz, 2016). To give a concrete example, in the current version of DBpedia's ontology (DBpedia 2015-10), the class *Canal* is classified as a sibling class of *River*, and both are defined as subclasses of *Stream*. This seems to be a rational classification at first glance since canals are usually channels of water. However, *Stream* is a subclass of *BodyOfWater* and *BodyOfWater* is a subclass of *NaturalPlace*. Due to the transitivity of the *rdfs:subClassOf* relationship, canals become natural places. However, this seems like an odd modeling choice as canals are defined as "an artificial waterway constructed to allow the passage of boats or ships inland or to convey water for irrigation" according to the Oxford dictionary. Words such as "artificial" and "constructed" make canals man-made features rather than natural place. This example indicates that top-down geo-ontologies may suffer from the issues such as modeling biases, oversights, and ontological commitments that do not well represent the real data needs.

Scrutinizing the geo-ontologies and making revisions manually on a regular basis are common solutions to such problems. But such methods are usually labor-intensive and create a gap between the geo-ontology and its corresponding Linked Dataset. In this research, we introduce initial results on a *Discrepancy Index* that helps geo-ontology engineers by detecting and quantifying potential issues using a series of data mining steps.

2 Proposed Method

Our approach consists of two parallel threads. The first thread comes from Linked Datasets that are transformed from unstructured data, such as Wikipedia pages. This thread focuses on the bottom-up

part. The second thread originates from the top-down geo-ontologies which are constructed manually by expert with their domain knowledge.

From a Linked Dataset, we select instances and properties concerning the specific classes in the top-down ontology. These instances and properties then act as input for our data-driven approaches. During the feature extraction, we focus on properties in each class. Properties in a Linked Dataset are analogous to attributes of different place types. The rationale is that similar place types share similar attributes while distinct place types have distinct attributes. For example, the place types *City* and *Town* are similar and they have similar properties such as *populationOf* and *totalAreaOf*. However, *City* and *Mountain* are very different from each other because a mountain can have a peak whereas a city usually does not. Based on this train of thought, we build a feature set that shows the relative frequency of each property in each class.

In pursuance of comparing the results of our data-driven approach, we also consider different variations of the feature set. We take into account four variables in our feature selection. They are *filler*, *specificity*, *literal* and *uniformity* (Table 1). All of them are boolean variables. The variable *filler* decides whether we use {property, object type} pairs or property alone to count the frequency. The variable *specificity* takes into consideration the hierarchical structure of object types. The variable *literal* acknowledges the fact that, in Resource Description Framework (RDF), object and literal have different typing schemes, namely object type and data type. The variable *uniformity* considers the cases in which the literal of the same property has different data types because the original Wikipedia page does not define a uniform data type for each infobox entry. For example, a city may have a property *population*, but the literal value for this property may be of type integer, double or even string. In a nutshell, *filler* and *literal* decide whether we incorporate object type and literal type into our feature extraction; *specificity* and *uniformity* deals with the granularity and accuracy of object type and data type respectively.

Table 1: Definition for four variables

	True	**False**
Filler	Include object types	Do not include object types
Specificity	Include only the most specific object types	Include all object types
Literal	Include literal types	Do not include literal types
Uniformity	Unify literal types	Do not unify literal types

Considering all the variables listed here, we have seven feature sets. This whole process of feature selection can be viewed as a decision making process, visualized in a decision tree shown in Figure 1. We make a boolean decision on one of the four variables described above at each internal node. Each internal node branches into two sub-nodes which are the outcomes of the two decisions based on each of the four variables in our feature selection process. The seven leaf nodes are the final outcomes of the decision tree, which are also the seven feature sets.

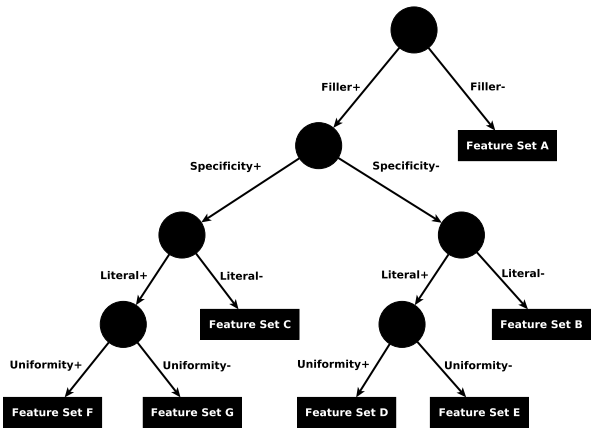

Figure 1: The tree structure of feature sets

In order to avoid the curse of dimensionality, we transform our raw feature space into a lower dimensional space using Multidimensional Scaling. We then use hierarchical clustering to obtain the hierarchies derived from the selected feature sets. In the next step, information content-based semantic similarity measures (Jiang and Conrath, 1997; Sánchez et al., 2011) are implemented to compute the pairwise similarity between each pair of classes in the derived as well as the original ontology hierarchy. The Mantel test (Mantel, 1967) is implemented to select the best representing feature set based on the correlation coefficient and p-value. The result shows that Feature Set E performs the best. In the end, we

use the semantic similarity results from the best feature set to calculate the *Discrepancy Index Matrix* $IndexMatrix = SimMatrix_{original} - SimMatrix_{derived}$. Individual *Discrepancy Index* can be found in this matrix.

The value range for the *Discrepancy Index* is $[-1, 1]$. If $IndexMatrix(i, j) > 0$, it implies that class i and j are less similar in the derived hierarchy; whereas if $IndexMatrix(i, j) < 0$, it tells us that class i and j are more similar in the derived geo-ontology. The value of $|IndexMatrix(i, j)|$ gives us the information about the extent to which the similarity in two hierarchies differ from each other. This *Discrepancy Index* is useful in assisting geo-ontology engineers to refine and further develop the geo-ontology in that it gives guidance on correcting the potential modeling biases or misclassifications.

3 Case Study: Are *Canals Natural Places*?

Among the various different types of geo-ontologies, we selected the DBpedia ontology for our case study. Previously, we discussed that *Canal* should not be classified as the subclass of *NaturalPlace* based on its definition. To further verify this, we can check if the DBpedia Linked Dataset supports our observation. Browsing through the DBpedia page of Panama Canal, we found properties such as *dbo:principalEngineer*, *dbp:dateUse* and *dbp:company* (see Figure 2). Intuitively, it is unreasonable for an instance of *NaturalPlace* to have principal engineers, to have the date of first use, and to be originally owned by a company. We can apply the *Discrepancy Index* to see if our approach can detect such a case.

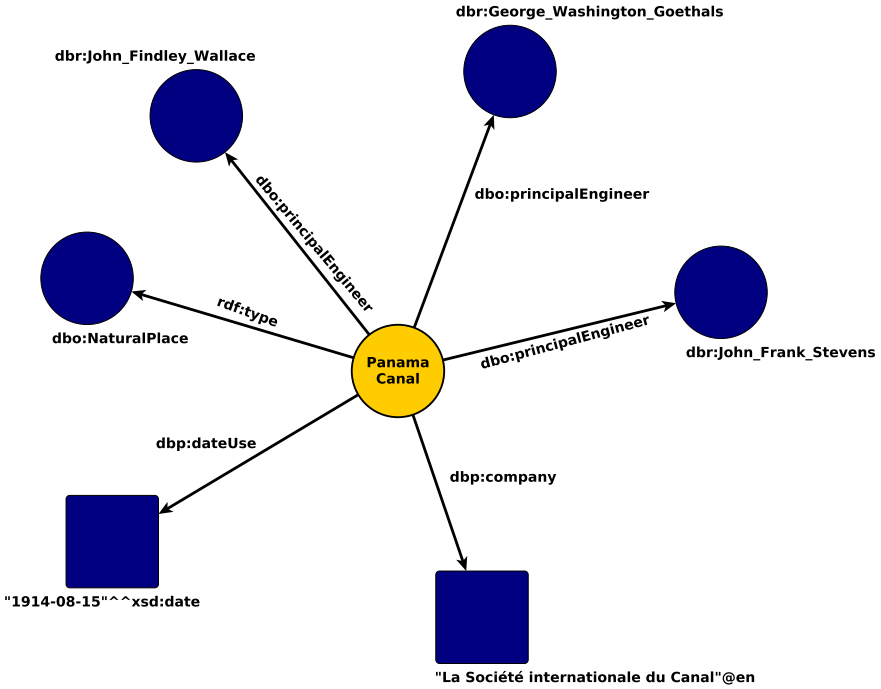

Figure 2: Panama Canal in DBpedia

We approach this by comparing the semantic similarity between *Canal* and its sibling classes, in this case, *River*. We would assume that since *Canal* and *River* are sibling classes, they are semantically similar to each other or at least more similar to each other than to classes in other branches of the place type hierarchy. We first plot the bar graphs for *Canal* and *River*. The bar chart uses information from the feature space after dimensionality reduction. The reasons we use the transformed feature space are two-fold. First, the original feature space contains numerous features which are hard to visualize and plot in a graph. Moreover, the original feature space is very sparse, making it difficult to analyze the results in a plotted graph. Second, some of the features are dependent on each other, making some of the information redundant and unreliable. The transformed feature space only contains 63 features and all of the features are independent from each other. Figure 3 and Figure 4 show the bar chart of these two classes. The x axis represents the statistical features for each class in the same order while the y axis is the value for each feature. From here, we can tell that *Canal* has a distinct pattern from *River*. However, the charts alone cannot quantify the difference between these classes. Moreover, it is also impossible to obtain the same kind of graphs using the original hierarchy in the ontology. Thus, these charts alone cannot hint at the direction of required geo-ontology refinements and the extent of the modeling biases.

Therefore, we can use the proposed *Discrepancy Index* to help us. After looking up the similarity matrices for the original hierarchy and the derived hierarchy, we find that the similarity between *Canal*

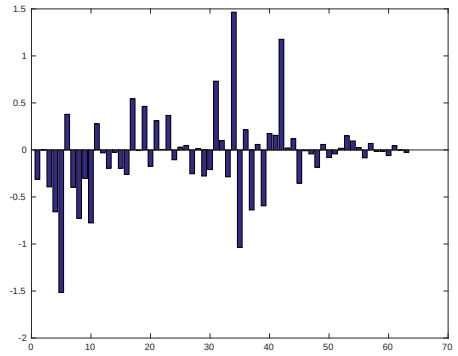

Figure 3: Feature values for *Canal* after MDS

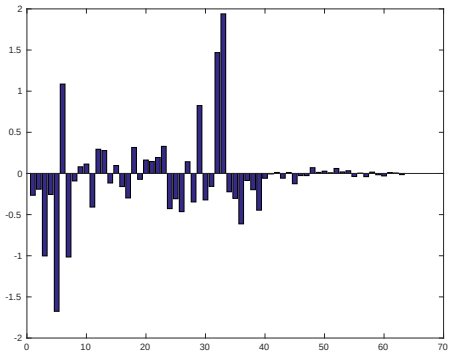

Figure 4: Feature vlaues for *River* after MDS

and *River* is 0.84 and 0.23 respectively. The *Discrepancy Index* for (*Canal*, *River*) is 0.61. This index implies that *Canal* is less similar to *River* in the derived hierarchy and leads to further investigation, which in this case is the result of modeling bias. This result corresponds to our observation in the DBpedia geo-ontology.

4 Conclusion and Future Work

Current geo-ontology engineering procedures often heavily depend on the knowledge of domain experts and a top-down style of engineering. The potential pitfall to this routine is that the resulting geo-ontologies may be biased and not representative of the data that will be semantically lifted using these ontologies. In this initial research we propose a data-driven approach that integrates geo-ontologies and Linked Dataset during the dynamic course of geo-ontology engineering and assist engineers in identifying and quantifying potential geo-ontology modeling bias via a *Discrepancy Index*. The initial case study suggests that the results returned by our method correspond to our observation, hinting at the usefulness of the *Discrepancy Index*.

This work can be extended in several aspects. First, this initial method can be extended into a systematic framework that can be applied to a variety of geo-ontologies and to guide engineers in understanding differences and similarities in their conceptualizations Janowicz et al. (2008) . Second, our experiment so far focuses on one particular ontology and dataset from DBpedia. With a wide range of availability of Linked Data and ontologies on the Web, we can test our approach using different data sources. Moreover, candidate solutions to the bias detected by the data-driven method can be developed in future work.

References

Hu, Y. and Janowicz, K. (2016). Enriching top-down geo-ontologies using bottom-up knowledge mined from linked data. In Onsrud, H. and Kuhn, W., editors, *Advancing Geographic Information Science: The Past and Next Twenty Years*, chapter 13, pages 183–198. GSDI Association Press.

Janowicz, K., Maue, P., Wilkes, M., Schade, S., Scherer, F., Braun, M., Dupke, S., and Kuhn, W. (2008). Similarity as a quality indicator in ontology engineering. In *Formal Ontology in Information Systems*, pages 92–105. IOS Press.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Sánchez, D., Batet, M., and Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297–303.