The relevance of measurement data in environmental ontology learning

Markus Stocker[1], Mauno Rönkkö[1],
Ferdinando Villa[2], and Mikko Kolehmainen[1]

[1] University of Eastern Finland,
P.O. Box 1627, Kuopio, Finland,
{markus.stocker,mauno.ronkko,mikko.kolehmainen}@uef.fi
[2] Basque Centre for Climate Change [BC3],
Alameda Urquijo 4-4, 48008 Bilbao, Spain,
ferdinando.villa@bc3research.org

Abstract. Ontology has become increasingly important to software systems. The aim of ontology learning is to ease one of the major problems in ontology engineering, i.e. the cost of ontology construction. Much of the effort within the ontology learning community has focused on learning from text collections. However, environmental domains often deal with numerical measurement data and, therefore, rely on methods and tools for learning beyond text. We discuss this characteristic using two relations of an ontology for lakes. Specifically, we learn a threshold value from numerical measurement data for ontological rules that classify lakes according to nutrient status. We describe our methodology, highlight the cyclical interaction between data mining and ontologies, and note that the numerical value for lake nutrient status is specific to a spatial and temporal context. The use case suggests that learning from numerical measurement data is a research area relevant to environmental software systems.

Keywords: Ontology, learning, rule-based reasoning, environmental data

1 Introduction

Ontology, defined as an explicit specification of a conceptualization [4], is a means to formally represent knowledge of a domain, meaning the concepts of some area of interest and relations that hold among them. Domains such as bioinformatics have used ontologies for over a decade [2, 1]. More recently ontologies have found applications in ecoinformatics [16] and environmental modelling [15]. With the development of ontologies it became clear that one of the major problems in ontology engineering is the often labour-intensive and time-consuming construction

[18]. Therefore, efforts have been on-going to automate the ontology acquisition, construction and maintenance processes [13].

Much of the effort within the ontology learning community has focused on learning from text collections, lexical databases, structural data, and usage data [18]. To the best of our knowledge, the state of the art in ontology learning mainly consists of several methods and tools to learn text entities such as words, concepts, relations, and noun hierarchies using machine learning and natural language processing [18, 13]. Learning beyond text is an open issue [18].

We investigate *environmental* ontology learning. Specifically, for the rule $p \rightarrow q$, meaning the implication between the antecedent p and the consequent q, we demonstrate the learning of a data value for an atom of p from sets of tuples with *numerical* elements obtained by measurement. Measurement is taken here to be the "process of empirical, objective assignment of numbers to the properties of objects and events of the real world in such a way as to describe them" [**?**].

Measurement is fundamental to environmental science: hence the relevance of numerical measurement data in environmental ontology learning. In environmental informatics, computational methods developed within disciplines such as data mining, machine learning and pattern recognition are routinely used to learn from data obtained by measurement, e.g. a linear regression model to forecast ambient ozone concentration. We demonstrate the application of such methods to environmental ontology learning. Further, we show that ontological knowledge can serve as a heuristic to guide the parametrization of data mining algorithms. Hence, we highlight an example of a data mining with ontology cycle [11] whereby ontological knowledge is used in data mining and the knowledge discovered from the resulting models is formalized and added to the ontology.

2 Materials and Methods

The environmental ontology used here is based on a taxonomy of lakes [17]. According to the trophic system for the classification of lakes, there are three main types of lakes, i.e. oligotrophic, eutrophic and heterotrophic [10, 14]. Several modifications have been proposed to this basic classification to account for our increased understanding of the lake ecosystem. The taxonomy of lakes adopted here extends the basic trophic system in that it evolves the naming convention to include the physico-chemical nature of water, the climatic zone, the type of lake basin, and the dominant class of organisms [17].

For the purpose here, we focus our attention on two properties for the physico-chemical nature of water, specifically the two ontology relations `richIn` and `poorIn` for the nutrient status of a lake with respect to nitrogen, phosphorus, and humus. Naturally, the question arises what being rich and poor in a nutrient for a lake exactly means and, thus, how to tell a lake is rich or poor in a given nutrient. We can refine the two relations with the suitable ranges for the concentrations [17].

In our implementation, we define two rules and use rule-based reasoning to infer the knowledge on whether an individual lake is rich or poor in a nutrient,

more accurately the nutrient status as measured by an individual lake monitoring station. Specifically, we consider the mean annual total nitrogen concentration. Thus, we learn the data (threshold) value `?y` of the atoms `lessThanOrEqual(?x, ?y)` and `greaterThan(?x, ?y)` for the rules

$$\texttt{totalNitrogen(?i, ?x)} \wedge \texttt{lessThanOrEqual(?x, ?y)} \rightarrow \texttt{poorIn(?i, Nitrogen)}$$

$$\texttt{totalNitrogen(?i, ?x)} \wedge \texttt{greaterThan(?x, ?y)} \rightarrow \texttt{richIn(?i, Nitrogen)}$$

where `?i, ?x, ?y` are variables. Informally, the rules state that an individual (lake monitoring station) `?i` with measured total nitrogen concentration `?x` $\leq$ `?y` is `poorIn` nitrogen. Conversely, an individual `?i` with measured total nitrogen concentration `?x` $>$ `?y` is `richIn` nitrogen. The rules are encoded in Jena[3] [3] (version 2.6.4) and the Jena general purpose rule engine is used for rule-based reasoning. Jena is a Java framework for building Semantic Web [7] applications.

Our aim is to learn the threshold value `?y`. For this purpose we use the k-means clustering algorithm [8] as implemented in WEKA[4] [5] (version 3.6.4) using data on the nutrient concentration of European lakes[5] (version 10) compiled by the European Environmental Agency (EEA). The two ontological relations `poorIn` and `richIn` suggest a binary classification of lakes with respect to nutrient status. This knowledge is used as a heuristic for the number of k-means clusters. Thus, we perform k-means clustering such that the unsupervised algorithm learns *two* centroids for the two-cluster separation of data on the mean annual total nitrogen concentration of lakes, typically for the lakes of a specific country as measured for a year. The values of the resulting k-means centroids represent a central tendency for the value of a lake `poorIn` and `richIn`, respectively. We define the threshold value `?y` to be the mean value for the two centroids. This modelling result is added as new knowledge to the ontology, specifically as knowledge about the two rules. Given an ontology for individual lakes we can, hence, use rule-based reasoning to infer new knowledge on lakes that are `poorIn` and those that are `richIn` nitrogen. Note that the choice of the mean for the two centroids as threshold value is for simplicity and may not be the most sensible as it may fall into one of the two clusters. Naturally, a different computation may be used.

For better data handling, we imported the EEA datasets for lake monitoring stations (3201 records) and for nutrients and organic matter in water (mean annual concentration for both total nitrogen and total phosphorus, 30866 records) into a PostgreSQL[6] database. We use the Resource Description Framework[7] (RDF) [9] language to represent information about lake monitoring stations, in

[3] `http://jena.sourceforge.net/`
[4] `http://www.cs.waikato.ac.nz/ml/weka/`
[5] `http://www.eea.europa.eu/data-and-maps/data/waterbase-lakes-6`
[6] `http://www.postgresql.org/`
[7] `http://www.w3.org/TR/rdf-primer/`

particular the corresponding lake name and the measured mean annual total nitrogen concentration. The Jena general purpose rule engine, customized with the learned rules, is used for rule-based reasoning. The SPARQL[8] [12] query language for RDF is used to query the resulting inference model for lakes `richIn` and lakes `poorIn` nitrogen.

Table 1. Centroid values (mg L^{-1}), threshold values `?y` (mg L^{-1}) and number of lake monitoring stations for lakes `poorIn` and `richIn` mean annual nitrogen concentration for 7 European countries in 2008.

	`poorIn`		`richIn`		
Country	Centroid	# stations	Centroid	# stations	`?y`
Denmark	0.57	15	1.52	4	1.04
Finland	0.39	150	0.88	53	0.63
Germany	0.72	24	3.84	4	2.28
Great Britain	0.36	53	4.51	3	2.43
Italy	0.82	127	2.46	9	1.64
Spain	0.78	137	8.36	12	4.57
Switzerland	0.66	6	1.76	4	1.20

3 Results

We learn the threshold value `?y` for the `poorIn` and `richIn` rules for the mean annual total nitrogen concentration (mg L^{-1}) as measured by 203 lake monitoring stations for Finnish lakes in 2008. There are a total of 203 measurements. The values of the two centroids as learned by WEKA using k-means are 0.39 and 0.88. They represent the central tendency for the value of `poorIn` and `richIn` mean annual nitrogen concentration for Finnish lakes in 2008, respectively. Thus, the threshold value `?y` is 0.63. SPARQL queries on a corresponding inference model return 150 lake monitoring stations for lakes `poorIn` and 53 `richIn` total nitrogen.

We perform a similar experiment for the threshold value for the mean annual nitrogen concentration as measured by 149 lake monitoring stations for Spanish lakes in 2008. There are a total of 149 measurements. The values of the two centroids are 0.78 and 8.36 for `poorIn` and `richIn`, respectively. Thus, the threshold value `?y` is 4.57. For Spain in 2008, there are 137 lake monitoring stations for lakes `poorIn` and 12 `richIn` total nitrogen.

Table 1 summarizes the central tendency for the value of a lake `poorIn` and `richIn` mean annual total nitrogen concentration for 7 European countries, in 2008. We also add the threshold value `?y` used in the corresponding rules. Further, the table shows the number of lake monitoring stations of lakes `poorIn` and lakes `richIn` total nitrogen. As the table highlights, the value of a lake

[8] `http://www.w3.org/TR/rdf-sparql-query/`

`poorIn` and `richIn` total nitrogen greatly varies between countries. In fact, for the listed countries the mean and standard deviation of the two centroids are 0.61 ± 0.18 and 3.33 ± 2.56, respectively.

Table 2. Centroid mean and standard deviation (mg L^{-1}) for lakes `poorIn` and `richIn` mean annual nitrogen concentration on data for the total number of years for 7 European countries. The table includes the number of years and the first year for which data is available for each country. The last year is 2008. Note that not all countries have data for all years.

			Centroid mean	
Country	# years	First	`poorIn`	`richIn`
Denmark	20	1989	0.65 ± 0.09	1.71 ± 0.36
Finland	33	1976	0.41 ± 0.02	0.95 ± 0.14
Germany	17	1991	0.69 ± 0.64	2.28 ± 1.91
Great Britain	14	1995	0.62 ± 0.24	3.87 ± 1.37
Italy	6	2003	0.53 ± 0.13	3.21 ± 3.88
Spain	1	2008	0.78 ± 0.00	8.36 ± 0.00
Switzerland	16	1993	0.78 ± 0.14	2.10 ± 0.48

Next, we analyse the variation of the value of a lake poor and rich in a nutrient over time. Table 2 summarizes the centroid mean and standard deviation for the central tendency and variation over time for lakes `poorIn` and `richIn` mean annual total nitrogen concentration for 7 European countries. For instance, for Finland Table 2 shows the mean and standard deviation for the centroids corresponding to `poorIn` (0.41 ± 0.02) and `richIn` (0.95 ± 0.14), for 33 years between 1976 and 2008. Figure 1 shows the variation for Finland over 33 years between 1976 and 2008. As expected from Table 2, there is considerable variation, in particular for the value of a lake `richIn` total nitrogen, for Finnish lakes over the time period.

4 Discussion

Given the rule $p \rightarrow q$, the main aim of this paper is to demonstrate the learning of a data value for an atom of the antecedent p from sets of tuples with *numerical* elements obtained by measurement. We have shown this using an environmental ontology for a taxonomy of lakes with the mean annual total nitrogen concentration, as measured by lake monitoring stations. We learned a threshold value for lakes poor and rich in total nitrogen as the data value for an atom of p. Given p with the rule atom `totalNitrogen(?i, ?x)` for the total nitrogen as measured by a lake monitoring station and the inequality rule atom, e.g. `lessThanOrEqual(?x, ?y)`, with the learned threshold value `?y`, we can apply rule-based reasoning to infer new knowledge on the two consequent q, `poorIn` and `richIn`. Hence, we demonstrate the relevance of numerical measurement data in environmental ontology learning.

Further, we have used ontological knowledge about the two-classes separation of lakes with respect to nutrient status as a heuristic to guide k-means in learning the centroids of the two classes. Thus, we describe an example of a cyclical interaction between data mining and ontologies [11].

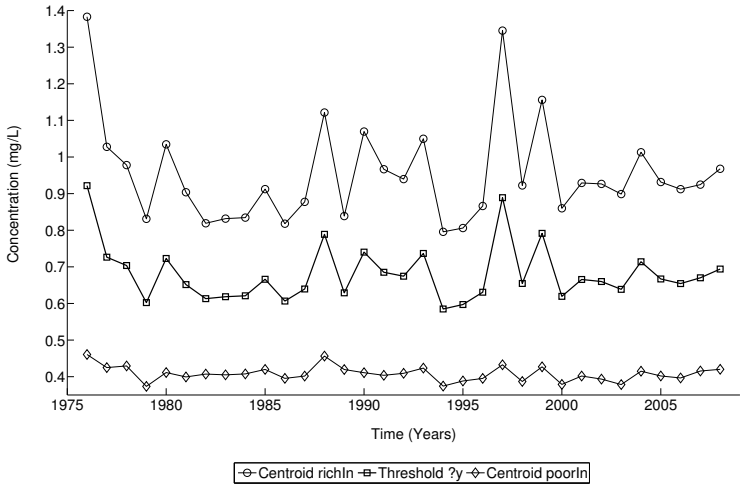

Fig. 1. Centroid values as central tendencies for lakes `poorIn` and `richIn` mean annual total nitrogen concentration and threshold value `?y` for Finland over 33 years between 1976 and 2008.

Our results show that whether a lake is poor (or rich) in total nitrogen is dependent on spatial context. As summarized in Table 1, the value for a lake poor or rich in total nitrogen clearly varies between countries. For instance, the central tendency of a Finnish lake rich in total nitrogen (0.88) is closer to the central tendency of a Spanish lake *poor* in total nitrogen (0.78) than to a Spanish lake rich in total nitrogen (8.36). Given the Finnish threshold value (0.63) for a Spanish lake with a mean annual total nitrogen concentration of $2.0\,\mathrm{mg\,L^{-1}}$ a rule-based reasoner would wrongly classify the Spanish lake as rich in total nitrogen.

Similarly, the results show that whether a lake is poor (or rich) in total nitrogen is dependent on temporal context. As summarized in Table 2, in particular by standard deviation, and shown in Figure 1, the value for a lake poor or rich in total nitrogen clearly varies over time. Thus, the threshold value for Finnish lakes for the year 1994 (0.58) is considerably different from the threshold value for the year 1997 (0.89).

Hence, environmental ontologies may be specific to spatial and temporal contexts. While this is unlikely to surprise the limnologist, we argue that environmental ontologies should reflect such spatial and temporal variation, and environmental ontology learning may provide methods and tools to automatically

adapt ontologies to spatial and temporal context using appropriate measurement data.

Moreover, the spatial and temporal variability raises the question why the value of a lake rich in nitrogen is significantly different between, say, Finland and Spain. Ontologies may provide a foundation for explanation services, in particular if ontological knowledge about the sources of nitrogen for lakes is available. For instance, a service may conclude that the spatial variability is (partially) explained by a different application of inorganic nitrogen fertilizers, properties of soil surrounding lakes that may affect the leaching of nitrogen, nitrogen fixation by cyanobacteria, or atmospheric deposition of nitrogen.

The use case discussed here is relatively straightforward, in particular with respect to the method used to learn the central tendency for the value of a lake poor or rich in a nutrient. While other clustering algorithms may be used for our aim of learning a threshold value, the chosen method reflects the rather simple, univariate, data. We may think of a use case whereby a decision tree for multivariate data is used to learn ontological rules, or to learn datatype restrictions for axioms of a knowledge base,[9] or to couple a trained neural network with an ontology, to classify a lake. Moreover, data-driven methods may help to uncover altogether new ontological classes or relations. For instance, cluster analysis may identify a set of lake classes different from that of the basic, or more advanced, trophic system and, therefore, affect the knowledge encoded in a corresponding ontology more profoundly.

Some of the techniques presented in this paper have been suggested elsewhere. Henson *et al.* [6] use a rule for high winds that states that a wind observation measurement greater or equal to 35 miles per hour is considered to be a high winds observation. Similarly to our example for the nutrient concentration of a lake, the authors use the rule for wind speed to infer new knowledge on high winds observations. Contrary to their example on wind, where the speed is known a priori, in our example we learn the meaning of `richIn` (and `poorIn`) from lake measurement data, i.e. from measurement of properties of real-world objects that exist in the domain modelled by the ontology.

5 Conclusions

We aimed at demonstrating the learning of ontological rules using numerical measurement data and clustering methods, specifically for an environmental ontology of lakes using k-means and the mean annual concentration of total nitrogen as measured by lake monitoring stations. Given the learned rules, we applied rule-based reasoning to infer new knowledge on the nutrient status of lakes. We described an example that shows the relevance of numerical measurement data in environmental ontology learning and the interaction between data mining and ontology engineering. The results of our experiments using the presented

[9] For instance, we may learn the interval for the basic ratio [17] datatype restriction `basicRatio some double[>= 0.0, < 1.2]` which is a property restriction in the terminological axiom that defines the ontological concept of *eutrophic lake*.

methodology on data for the lakes of a number of European countries highlight the expected spatial and temporal dependency of the numerical meaning of lake nutrient status, a characteristic that may justify further attention in the field of environmental ontology learning.

In our future work, we intend to develop a software prototype that implements the core ideas of the methodology presented in this paper. In particular, we envision the development of an ontology to describe learning tasks for the software to perform. Such an ontology may integrate the source and description of numerical data, the target ontology and what ought to be learned about it, as well as the method used for learning. We think such a software may support the learning of more complex environmental ontologies and, ultimately, lead to methodological generalizations.

Acknowledgements. We wish to thank Dr. Eila Torvinen, Ph.D., university researcher in the Environmental Microbiology Research Group at the University of Eastern Finland, and Dr. Bijan Parsia, lecturer in the School of Computer Science at the University of Manchester (UK), for their expertise, critique, and suggestions in numerous discussions. Further, we wish to thank the European Environmental Agency for providing open access to data, a service without which this work would not have been possible.

References

1. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G.: Gene ontology: Tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
2. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL. Nucleic Acids Res 31, 365–370 (2003)
3. Carroll, J.J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K.: Jena: Implementing the Semantic Web Recommendations. Tech. Rep. HPL-2003-146, HP Laboratories, Bristol, UK (2003)
4. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. In: SIGKDD Explorations. vol. 11 (2009)
6. Henson, C.A., Pschorr, J.K., Sheth, A.P., Thirunarayan, K.: SemSOS: Semantic Sensor Observation Service. In: Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems (CTS 2009). Baltimore, MD (May 2009)
7. Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (2001)
8. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)

9. Manola, F., Miller, E.: RDF Primer. Tech. Rep. W3C Recommendation, W3C (2004)
10. Naumann, E.: Nagra synpunker angaende planktons okologi. Med sarskild hansyn till fytoplankton. Svensk bot. Tidskr. 13, 129–158 (1919)
11. Nigro, H.O., Císaro, S.E.G., Xodo, D.H.: Data mining with ontologies: Implementations, findings, and frameworks. Information Science Reference (an imprint of IGI Global) (2008)
12. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. Tech. Rep. W3C Recommendation, W3C (2008)
13. Shamsfard, M., Barforoush, A.: The state of the art in ontology learning: A framework for comparison. Knowledge Engineering Review 18(4), 293–316 (2003)
14. Thienemann, A.: Physikalische und chemische Untersuchungen in den Maaren der Eifel. Verh. Naturh. Ver. preuss. Rheinl. u. Westfalens 71, 281–389 (1915)
15. Villa, F., Athanasiadis, I., Rizzoli, A.: Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. Environmental Modelling and Software 24(5), 577–587 (2009)
16. Williams, R., Martinez, N., Golbeck, J.: Ontologies for ecoinformatics. Web Semantics 4(4), 237–242 (2006)
17. Zafar, A.: Taxonomy of lakes. Hydrobiologia 13(3), 287–299 (1959)
18. Zhou, L.: Ontology learning: State of the art and open issues. Information Technology and Management 8(3), 241–252 (2007)