Comparing Alternative Data-Driven Ontological Vistas of Natural History

Marieke van Erp, Piroska Lendvai, and Antal van den Bosch

Tilburg centre for Creative Computing, Tilburg University, The Netherlands
{M.G.J.vanErp, P.Lendvai, Antal.vdnBosch}@uvt.nl

Abstract

Traditionally, domain ontologies are created manually, based on human experts' views on the classes and relations of the domain at hand. We present ongoing work on two approaches to the automatic construction of ontologies from a flat database of records, and compare them to a manually constructed ontology. The latter CIDOC-CRM ontology focusses on the organisation of classes and relations. In contrast, the first automatic method, based on machine learning, focuses on the mutual predictiveness between classes, while the second automatic method, created with the aid of Wikipedia, stresses meaningful relations between classes. The three ontologies show little overlap; their differences illustrate that a different focus during ontology construction can lead to radically different ontologies. We discuss the implications of these differences, and argue that the two alternative ontologies may be useful in higher-level information systems such as search engines.

1 Introduction

Ontologies are explicit conceptualisations of domains. A vast amount of work on ontologies in the knowledge representation field has focussed on their use in facilitating knowledge sharing between knowledge-based systems, and in the interaction between such systems and human users [3]. Ontologies can for instance offer help in visualising the domain for users, and hence improve their understanding of the information, or they can be employed to enhance searching in domain data through for instance query expansion or faceted navigation.

It is conceivable to have different ontologies for a single domain. Although developers of traditional ontologies tend to stress that "true" ontologies are function-independent, in a more practical sense it is possible to design ontologies with a particular functionality in mind, such as an embedding in an information system. This may influence design choices. For certain tasks, a more formal and elaborate ontology is required, whereas for other tasks a simpler conceptualisation of the domain that only contains the most important classes and relations may suffice. Such considerations may influence choices when designing an ontology, as ontology construction is an expensive task, traditionally requiring knowledge from and formalisation by or with domain experts.

In the past decade, an increasing amount of work is invested in the development of support systems for automatic or semi-automatic ontology construction, with workshops devoted to this topic at several AI conferences such as ECAI and IJCAI [1]. In this study, three ontologies for a single domain are presented. The scenario is that at the outset we have a database of records, each describing an instance of an object – in our case study, a zoological specimen in a natural history collection, described by several textual fields. The database column labels can serve as starting points for naming the class nodes in our ontology. The task then is to find out how these classes are related to each other; we let two data-driven methods induce these relations. As a gold standard for comparing our two automatic ontology construction methods, we also have a manually designed ontology for the domain.

2 Three Ontologies

The database used as a starting point in this paper describes key characteristics of reptile and amphibian (R&A) specimens present in the collection of the Dutch National Museum for Natural History[1], using mostly textual database fields. It is constructed manually, and contains 16,870 records in 39 columns. Most values are limited to one or a few words, some fields contain longer stretches of text, for instance describing the climatological circumstances or the location at which a specimen was found.

2.1 A Hierarchical Ontology

As a baseline, an ontology was manually constructed following the CIDOC-CRM conceptual model standards [2] (henceforth: CIDOC). It is relatively straightforward to match each column (representing a salient domain concept) and its relevant relations from the R&A database to a class in CIDOC. The prime goal of the CIDOC reference ontology is to offer a framework to model the *organisation* of processes and entities within a cultural heritage collection. This goal leads to a richness in hierarchical relations, expressing mainly hyponymy and meronymy relations. In Figure 2.3, these relations are indicated by the uninterrupted lines.

2.2 A Mutual Predictiveness Ontology

The second ontology is based on the application of machine learning methods to the R&A database. It aims to reflect the predictability of values in one database column on the basis of values in other columns. In ontological terms: knowing the values of certain fields in one instance, the values of certain other fields may be predicted accurately. Indeed, we show by performing machine learning experiments that in our database certain columns are conditionally dependent on each other. For instance, if the "Province" field of a record has the value *West Java*, most machine learning methods can be trained to induce that the value in the "Country" field must be *Indonesia* given enough training samples

[1] http://www.naturalis.nl

of database records. Such conditional dependencies can be directly used for our current goal: establishing relations between classes. When a machine learning algorithm (such as a machine learning algorithm adopting an explicit feature selection preprocessing step) actively selects a database column to predict the values of another column, we assume that in the ontology the class nodes belonging to the two database fields are connected by a directed "predictiveness" relation. In Figure 2.3, the dotted lines represent the relations between a class and its single-most predictive class.

2.3 A Hybrid Ontology

The second data-driven ontology proposed here again utilises the R&A database, as well as a human-made semantic network resource, in order to look for possible relations between the classes. The database is a handy starting point, as each record is already a structured piece of information carrying instances of paired values. These pairs are subsequently looked up in the external semantic network resource, to verify whether this resource knows the particular relation between this pair – which may in turn be a good candidate label for the relation between the pairs' classes.

To this purpose we chose to use the online encyclopaedia Wikipedia[2]. Wikipedia's link structure can be considered a semantic network, as the guidelines for contributors state that links from the page of one concept to another should only be added when they are meaningfully related [4]. We find candidate relation labels between database columns by discovering relations, i.e. linguistic predicates between values from these columns co-occurring within the limits of a sentence, given that their Wikipedia pages are linked. The obtained verb phrases are ranked by frequency and annotated by human evaluators. In Figure 2.3 the relations in this hybrid ontology are indicated by the dashed lines. For the sake of clarity not all concepts within the domain are shown in the graph and relation labels are also excluded.

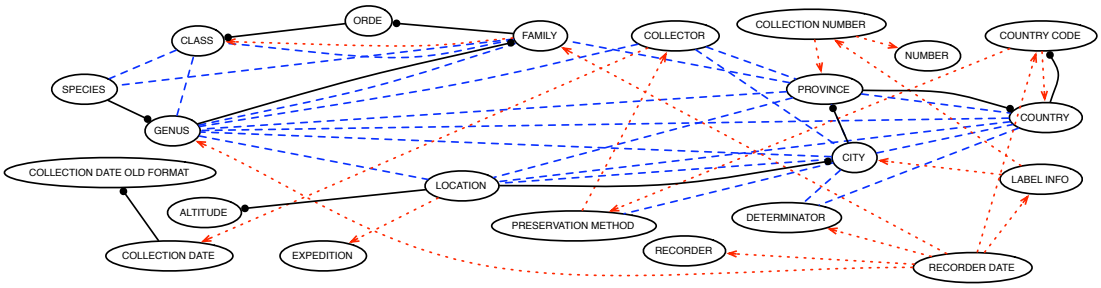

Figure 1: Relations from hierarchical, data-driven and hybrid ontologies

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

3 Discussion

The three ontologies presented in the previous section are remarkably different from each other; their overlap is minimal. This can only be attributed to the different building principles of the three ontologies.

In the machine-learning-based ontology, a relation signifies a conditional dependency relation between two classes. Interestingly, this method uncovers relations between classes of radically different entity types (such as between collectors and locations) that are yet meaningful in the domain. Conditional dependence can also be a guideline for data storage, as it indicates which information is redundant, and can thus be compressed or stored optimally.

The hybrid ontology offers a middle ground between the machine learning and CIDOC ontologies. It is created via analysing human-generated content in an external semantic resource, namely Wikipedia. The obtained relations originate from a pool of possible rankings by human judges, therefore we argue that this ontology represents relations in the natural history domain that are fairly accepted. Compared to the hybrid ontology, the CIDOC ontology is rather sparse; for instance, between the biological taxon concepts it strictly encodes the hypernym relations between parent and child nodes, whereas the hybrid ontology tends to link everything that is somehow related according to encyclopaedic evidence.

To conclude, we believe the hybrid approach is still crude, but it does produce possible links between domain concepts attested in an external encyclopeadic resource, while requiring little effort in development. We also believe that conditional dependence, as detectable through machine learning, should be considered as a ground for establishing relations between concepts. While the final decision should be left to human experts, both methods may serve as useful ontology expansion proposal methods.

References

[1] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005.

[2] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. Definition of the cidoc conceptual reference model. Technical report, ICOM/CIDOC CRM Special Interest Group, 2008.

[3] Thomas R. Gruber. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, chapter Toward Principles for the Design of Ontologies used for knowledge sharing, pages 907–928. Kluwer Academic Publishers, 1995.

[4] Jaap Kamps and Marijn Koolen. The importance of link evidence in Wikipedia. In Craig Macdonald et al, editor, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282, Heidelberg, 2008. Springer Verlag.