From Mappings to Modules: Using Mappings to Identify Domain-Specific Modules in Large Ontologies

Amir Ghazvinian, Natalya F. Noy, Mark A. Musen

Stanford University, Stanford, CA 94305, US
{amirg,noy,musen}@stanford.edu

Using Mappings to Identify Modules. Ontology modularization is an active area of research in the Semantic Web community [3]. With the emergence and wider use of very large ontologies, in particular in fields such as biomedicine, more and more developers need to extract meaningful modules of these ontologies to use in their applications. Researchers have also noted that many ontology-maintenance tasks would be simplified if we could extract modules from ontologies. These tasks include ontology matching: If we can separate ontologies into modules, we can simplify and improve ontology matching. We study a complementary problem: Can we use existing mappings between ontologies to facilitate modularization?

Methods. Figure 1 illustrates our method: First, we generate mappings from a source ontology to the target ontology that we wish to modularize. Next, we cluster mappings within the target ontology. Finally, we use mapping clusters to identify modules within an ontology.

Validation and Analysis. We validate and analyze our approach by applying our methods to identify modules for NCI Thesaurus [1] and SNOMED-CT [2], two popular and large biomedical ontologies. As domain-specific ontologies for modularization, we used 141 ontologies in BioPortal.[1] Our process extracted 71 modules for NCI Thesaurus and 68 modules for SNOMED-CT. We examined modules and their representative terms in order to understand the types of modules that our algorithm creates and to determine whether or not these modules are likely to be useful in an application setting. Figure 2 shows an example, a module of NCI Thesaurus that is relevant to electrocardiograms (EKG) using the Electrocardiography Ontology. The module consists of 61 classes, representative samples of which are shown in the figure. Of the 61 classes in this module, 41 (67%) are mapping targets.

Discussion and Conclusions. Our approach uses mappings between ontologies in order to extract domain-specific modules from large ontologies based on their mappings to smaller ontologies. In our experiments with NCI Thesaurus and SNOMED-CT, using the ontologies from BioPortal as the sources for mappings, we have identified a number of useful modules. We found that one of the key hurdles that we must overcome, is to find a way to determine how good a particular module is. Indeed, the same problem is true for most modularization approaches [3]: many authors discuss computational properties of their modules, but do not evaluate how useful these modules are to users. In our case, the requirements for extraction are driven by domain coverage of the module rather than by its computational or structural properties. Thus, the problem of evaluating whether the module satisfies the user requirements is similar to the problem of ontology

[1] http://bioportal.bioontology.org

evaluation in general: how do we know that an ontology is useful for a specific class of applications? We plan to submit the modules that we have identified to BioPortal to enable the user community to use the modules in their applications, review them and comment on them. However, our initial evidence, which we present in this paper, indicates that our approach can indeed find interesting domain-specific modules.

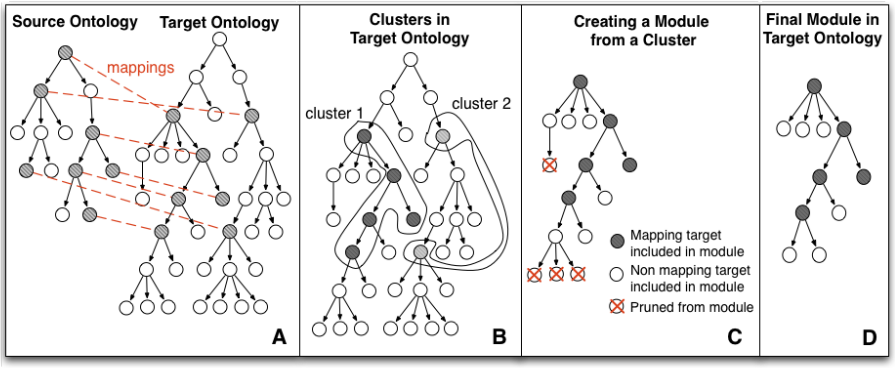
Source Ontology
Target Ontology
mappings
A
Clusters in Target Ontology
cluster 1
cluster 2
B
Creating a Module from a Cluster
Mapping target included in module
Non mapping target included in module
Pruned from module
C
Final Module in Target Ontology
D

Fig. 1. The process of identifying modules using mappings between ontologies. A: mappings between a source ontology and a modularization target. B: two clusters returned by clustering the mappings. One cluster is light gray in color while the other is dark gray. When determining a module based on these clusters, we discard the light gray cluster since the mapping targets within that cluster are too sparse. C: the process of pruning the ontology subtree for the remaining cluster, which we use to create the module. We begin at each leaf and traverse the tree toward the root, removing all classes that are not mapping targets or direct children of mapping targets. Once we reach such a class, we stop pruning along that branch. D: the final module.

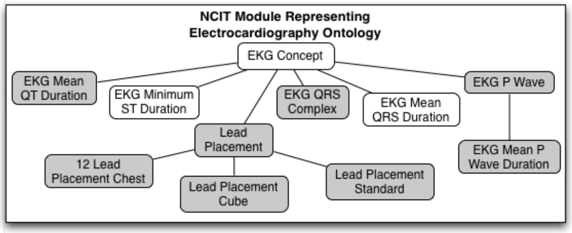

Fig. 2. A portion of the module that we identified within NCI Thesaurus that represents the domain of the Electrocardiography Ontology. The classes in gray represent mapping targets and classes in white represent classes that were not mapping targets, but are included in the module through our algorithm.

References

1. N. Sioutos, S. de Coronado, M. Haber, F. Hartel, W. Shaiu, and L. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.
2. K. Spackman, editor. *SMOMED RT: Systematized Nomenclature of Medicine, Reference Terminology*. College of American Pathologists, Northfield, IL, 2000.
3. H. Stuckenschmidt, C. Parent, and S. Spaccapietra, editors. *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*. Springer-Verlag, Berlin, 2009.