Merging DAML+OIL bio-ontologies

Patrick Lambrix, Anna Edberg, Carolyn Manis, He Tan
Institutionen för datavetenskap, Linköpings universitet, Sweden
Email: patla@ida.liu.se

Abstract

Ontologies are being used nowadays in many areas. Within the bioinformatics area there are a number of bio-ontologies that cover different aspects in molecular biology. Many of these ontologies contain overlapping information. In applications using multiple ontologies it is therefore of interest to be able to merge these ontologies. In this paper we describe a prototype implementation of an ontology merging tool for DAML+OIL ontologies. The tool generates suggestions for merging roles and concepts and for is-a relationships between concepts. We also evaluate the quality of the suggestions using well-known bio-ontologies.

1 Introduction

Ontologies define the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary. They are being used nowadays in many areas, including bioinformatics, for several reasons. Ontologies are used for communication between people and organizations by providing a common terminology over a domain. They provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information. Further, they can also be used as a basis for integration of information sources and as a query model for information sources (e.g. [7]).

Within the bioinformatics area there are a number of bio-ontologies, each with their own focus, that cover different aspects in molecular biology such as molecular function and cell signaling. Many of these ontologies contain overlapping information. For instance, a protein can be involved in both cell signaling and other biological processes. In applications using ontologies it is therefore of interest to be able to use multiple ontologies. However, to obtain the best results, we need to know how the ontologies overlap and align them or merge them into a new ontology. Another reason for merging ontologies is that it allows for the creation of ontologies that can later be composed into larger ontologies. Also, companies may want to use de facto standard ontologies and merge them with company-specific ontologies. There exist a number of tools for merging ontologies that have been used outside the bioinformatics area. In [6] we evaluated how well two well-known tools, Protégé-2000 with PROMPT [10] and Chimaera [9], perform for merging bio-ontologies. The ontologies we used for testing were Gene Ontology [3] ontologies (GO), which have become a de facto standard and Signal-Ontology [12] (SigO). The overlap between these ontologies gives interesting test cases.

In this paper we describe a prototype of the OntologyMerge system, that was developed taking the results of [6] into account. In its current implementation it helps users merge DAML+OIL (http://www.w3.org/TR/daml+oil-reference) ontologies based on string matching and description logic reasoning. DAML+OIL, which can be seen as an expressive description logic (e.g. [5]), has already gained acceptance as ontology representation language in the bio-ontology domain. This can be seen by the facts that the Bio-Ontology Consortium has specified DAML+OIL as their ontology exchange language, and that the GO ontologies are being migrated to DAML+OIL in the GONG project. In the paper we also describe a preliminary evaluation of the quality of the suggestions using GO and SigO.

2 System

The current implementation of OntologyMerge is a web-based system that helps a user to merge two DAML+OIL ontologies into a new DAML+OIL ontology with unique names for terms (currently concepts or roles). The OntologyMerge system provides a user with two modes: the manual mode and the suggestion mode. In the manual mode the user decides herself which terms in the two ontologies should be merged or should be related via an is-a relation. The system performs the actual merging and computes the consequences of the merge. In the suggestion mode the system generates suggestions for merging or for creating is-a relationships. This is done in three steps. First, suggestions for roles are generated, then suggestions for merging concepts and finally suggestions for is-a relationships between concepts. Each step should be finished before to the next step is started. Within each step the user can accept or reject the suggestions. When several terms from the second ontology are suggested for a term in the first ontology, the different suggestions are presented to the user at the same time. Upon acceptance of a suggestion, the system performs the actual merging or includes the is-a relationship, computes the consequences and makes the additional changes that follow from the operation. Upon rejection of a suggestion, it is checked that the terms in the rejected suggestion do not have the same name. If so, one of the terms needs to be renamed. In each case the suggestion list is updated. At each point in time the user receives information on which operations have been performed and how many suggestions there are left for the step. After all suggestions are processed, the system copies the terms of the original ontologies that are not merged to the new ontology.

The suggestions are generated based on string matching. We are currently evaluating different approaches. As a first step we have implemented different matching algorithms for the different steps. For the first step (roles) the system generates a suggestion when the names of the terms in the different ontologies are the same or when one is a prefix or a suffix of the other. For the second step (merging of concepts) a suggestion is generated based on an N-gram fuzzy string matching algorithm that is applied to the concept names. Here we also make explicit use of the fact that many bio-ontologies define a 'synonym' role and thus check whether synonyms match the concept names. Finally, in the third step (is-a relationships between concepts) the system checks whether a concept name in one ontology contains all the words of a concept name in the other ontology.

Tool	Case	Suggestions	Correct	Missing	Recall	Precision
PROMPT	B	3	3	2	0.6	1
Chimaera	B	16	4	1	0.8	0.25
OntologyMerge	B	5	5	0	1	1
PROMPT	ID	4	4	5	0.44	1
Chimaera	ID	10	4	5	0.44	0.4
OntologyMerge	ID	11	7	2	0.78	0.63

Table 1: Quality of suggestions

The OntologyMerge system uses further the FaCT [4] system to provide a number of reasoning tasks. (Our implementation is partly based on OilEd [1] ideas and implementation.) The user can ask the system to check whether the new ontology is consistent. She can receive information about unsatisfiable concepts and cycles in the ontology. The user can also receive an updated DAML+OIL ontology that is generated from the FaCT representation and therefore this ontology contains less redundancy as well as explicit statements for derived relationships.

3 Evaluation

We have performed an evaluation of the quality of the suggestions that our system generates. We compare the results with the results for Protégé-2000 with PROMPT and Chimaera as reported in [6]. Both systems give users suggestions for merging and when an operation is chosen, the systems perform additional changes that follow from that operation. Chimaera also generates a taxonomy resolution list where it suggests taxonomy areas that are candidates for reorganization. We note that for Protégé-2000 with PROMPT and Chimaera we have used the systems as they are provided in the basic distribution although more advanced variants exist.

Test Ontologies In our tests two chosen 'cases' from GO and SigO were used. Each case consists of one part of SigO and one part of GO. Each case was chosen in such a way that there was an overlap between the GO part and the SigO part. The first case, *behavior*, contains approximately 60 terms from GO and approximately 10 terms from SigO. The second case, *immune defense*, contains approximately 70 terms from GO and approximately 15 terms from SigO. A complete listing of the terms that are included can be found in [2]. We used more terms from GO than from SigO because the granularity of GO is higher than the granularity of SigO for these topics.

Evaluation Results In table 1 we show the results regarding the quality of the generated suggestions in terms of precision and recall for the cases *behavior* (B) and *immune defense* (ID), respectively. Precision measures how many of the suggestions are relevant while recall measures how many of the relevant suggestions the system actually proposed. OntologyMerge performs perfect with respect to quality of suggestions on the *behavior* case and good on the *immune defense* case. (If we would not have taken the synonyms into account, then we would have had the same results for the *behavior* case while for the *immune defense* case the precision would have been 0.66 and the recall 0.6.) Let us have a more detailed look at the *immune defense* case. Table 2 shows the suggestions created by the system. The first column describes

	Gene Ontology	Signal Ontology	
M	synonym	Synonym	C
M	immune response	Immune Response (synonym:immune response)	C
M	B-cell activation (synonym:B-cell proliferation)	B Cell Activation	C
M	B-cell activation (synonym:B-cell proliferation)	T Cell Activation	W
M	T-cell activation (synonym:T-cell proliferation)	T Cell Activation	C
M	T-cell activation (synonym:T-cell proliferation)	B Cell Activation	W
M	complement activation	Complement Signaling (synonym:complement activation)	C
M	antigen presentation	Antigen Processing and Presentation	W
M	antigen processing	Antigen Processing and Presentation	W
R	antigen presentation	Antigen Processing and Presentation	C
R	antigen processing	Antigen Processing and Presentation	C
R	*inflammatory response*	*Inflammation*	NF
R	*activation of natural killer cell activity*	*Natural Killer Cell Response*	NF

Table 2: Suggestions for *immune defense*

whether the suggestion was for a merge (M) or is-a relationship (R). The last column describes whether the suggestion was correct (C), wrong (W) or not found by the system (NF). In this case, the two missing suggestions are for is-a relationships between concepts. They might have been found by using stemming and more flexible sub-string matching. The wrong suggestions have to do with similar concepts where the matching was too flexible. We note, however, that the user receives multiple suggestions simultaneously if a concept from the first ontology is involved in multiple suggestions. For instance, (B-cell activation, B Cell Activation) and (B-cell activation, T Cell Activation) are shown simultaneously. We also note that when the user has decided on merging, the suggestion list is updated and therefore not all the wrong suggestions will actually be shown to the user.

4 Conclusion

We have described a merging tool for DAML+OIL ontologies and evaluated the system in terms of precision and recall for two well-known bio-ontologies. We also note that GO and SigO currently do not require the full expressivity of DAML+OIL. Therefore, the full description logic reasoning capabilities of the tool may not have been exploited yet in this test.

The system still needs to be evaluated with respect to other factors as described in [6] and [8] and such an evaluation is ongoing [13]. For instance, we already know that the user interfaces of Protégé-2000 and Chimaera have a more elaborate visualization of the ontologies than our current implementation. Also, these systems allow for many different input and output formats and are more full-fledged ontology engineering tools than our system.

We also want to evaluate different ways of providing suggestions to the user. We will experiment with different matching algorithms as well as with more advanced strategies such as, for instance, used in Anchor-PROMPT [11]. Finally, we will make it possible to interleave the manual and the suggestion modes of the system.

Acknowledgements We thank Vaida Jakonienė and Bo Servenius for comments on the system. We also acknowledge the financial support of the Center for Industrial Information Technology.

References

[1] S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens. OilEd: a reason-able ontology editor for the semantic web. *Proceedings of the Joint German Austrian Conference on Artificial Intelligence*, LNAI 2174, pp 396-408, 2001.

[2] A. Edberg. *Ontology merging in bioinformatics.* M.Sc. thesis LiTH-IDA-Ex-02/62, Linköpings universitet, Sweden, 2002. In Swedish.

[3] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25-29, 2000. http://www.geneontology.org/

[4] I. Horrocks. FaCT and iFaCT. *Proceedings of DL1999*, pp 133-135, 1999.

[5] I. Horrocks and U. Sattler. Ontology reasoning in the SHOQ(D) description logic. *Proceedings of IJCAI2001*, pp 199-204, 2001.

[6] P. Lambrix and A. Edberg. Evaluation of ontology merging tools in bioinformatics. *Proceedings of the Pacific Symposium on Biocomputing - PSB03*, pp 589-600, 2003.

[7] P. Lambrix and V. Jakonienė. Towards transparent access to multiple biological databanks. *Proceedings of the First Asia-Pacific Bioinformatics Conference*, pp 53-60, 2003.

[8] P. Lambrix, M. Habbouche, and M. Pérez. Evaluation of ontology engineering tools for bioinformatics. Accepted for publication in the *Bioinformatics* journal.

[9] D. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. *Proceedings of KR2000*, pp 483-493, 2000. Chimaera: http://www.ksl.stanford.edu/software/chimaera/

[10] N. Noy and M. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. *Proceedings of AAAI2000*, pp 450-455, 2000. Protégé: http://protege.stanford.edu/index.html, PROMPT: http://protege.stanford.edu/plugins/prompt/prompt.html

[11] N. Noy and M. Musen. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. *Proceedings of the IJCAI2001 Workshop on Ontologies and Information Sharing*, 2001.

[12] Signal-Ontology: http://ontology.ims.u-tokyo.ac.jp/signalontology/

[13] H. Tan. *Merging DAML+OIL Bio-ontologies.* M.Sc. thesis LiTH-IDA-Ex-03/51–SE, Linköpings universitet, Sweden, 2003. forthcoming.