Structural techniques for alignment of structurally dissymmetric taxonomies

Chantal Reynaud, Brigitte Safar, Hassen Kefi
LRI-PCRI Batiment 490 Université Paris-Sud
91405 Orsay Cedex France
firstname.lastname@lri.fr

ABSTRACT
This paper deals with taxonomy alignment and presents the structural techniques of an alignment method suitable with a dissymmetry in the structure of the mapped taxonomies. The aim is to allow a uniform access to documents belonging to a same application domain, assuming retrieval of documents is supported by taxonomies. We applied our method to various taxonomies using our prototype TaxoMap.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms
Design, Algorithms, Experimentation

Keywords
Taxonomy, Alignment, Mapping, Unified access

1. INTRODUCTION
Our work focuses on taxonomy alignment techniques. Indeed, we assume that description of content of most information systems is often based on very simple ontologies reduced for the present to classification structures, i.e. taxonomies. Moreover, we suppose that the structures of the taxonomies that we align are heterogeneous and dissymmetric, one taxonomy being deep whereas the other one is flat. Such a situation can be encountered for example when we try to access additional resources with very simple classification structures describing the domain concepts from a Web portal having its own query interface based on a hierarchically well-structured taxonomy. In this context, the approaches that rely on OWL data representations exploiting all the ontology language features don't apply [3]. Similarity of two entities cannot be identified based on their properties or on the status of their parents and siblings because this information is not available. To find mapping candidates between structurally dissymetric taxonomies, we can only use the following available data: labels of concepts in both taxonomies, the structure of the deeper taxonomy and external resources such as WordNet.
This paper described two structural techniques designed to make best use of the characteristics of the taxonomies: very specialized taxonomies with only subclass links, concepts with labels which are expressions composed of a lot of words, words common to a lot of labels. These techniques have been evaluated on real-world taxonomies and on test ones extracted from a repository about ontology matching [5]. Experiments showed that the proposed techniques give very relevant mappings when the aligned taxonomies have the same characteristics as those having motivated our approach.

Posters and Demos of the 15th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2006 Podebrady, Czech Republic, 2nd-6th October, 2006

2. THE ALIGNMENT PROCESS
For us, a taxonomy is a pair $(C,\ H_C)$ consisting of a set of concepts C arranged in a subsumption hierarchy H_C. A concept is only defined by two elements: a label and subclass relationships. The label is a name (a string) that describes entities in natural language and that can be an expression composed of several words. Subclass relationships establish links with other concepts. It is the single semantic association used in the hierarchy.
Given two structurally dissymmetric taxonomies, our objective is to map the concepts of the less structured one, the source taxonomy T_{Source}, with concepts of the more structured one, the target taxonomy T_{Target}.The alignment process is oriented from T_{Source} to T_{Target}. It aims at finding one-to-one mappings which are relations of two kinds: equivalence ($isEq$) and subclass (isA). So, for each concept c_S in T_{Source}, we try to find a corresponding concept c_T in T_{Target} linked to c_S with an equivalence or a subclass relation.

3. THE ALIGNMENT TECHNIQUES

3.1 General view
Alignment is based on Lin's similarity measure [1], computed between each concept c_S in T_{Source} and all the concepts of T_{Target}. This measure compares strings and has been adapted to take into account the importance of the words inside the expressions. Various techniques are applied in sequence to make the overall alignment process the most efficient as possible. For each technique, the objective is to select the best concept in T_{Target} among a lot of mapping candidates. This best concept is not necessarily the concept with the highest similarity measure. We classify the found mappings into two groups according to their relevance: likely

mappings and potential mappings to be confirmed.

Algorithm 1: Alignment process
$TaxoMap(T_{Source}, T_{Target})$
1. **For each** $c_S \in T_{Source}$ **do**
2. **For each** $c_T \in T_{Target}$ **do** $Sim_{LinLike}(c_S, c_T)$
3. $MC \leftarrow$ MappingCandidates(c_S)
4. **If** LikelyMapping(c_S, MC) **then** stop
5. **Else** PotentialMapping(c_S, MC)

Terminological techniques are executed first. In default of place, they will not be detailed here. Being based on the richness of the labels of the concepts, they provide the most likely mappings (cf. Alg.1). However a lot of mappings are not found. So we propose to complete these first techniques with two structural ones suited to our work context, deriving interesting but less certain (potential) mappings. So, a user evaluation of these new mappings is necessary.

3.2 Exploiting structural features

The two structural techniques that we proposed are complementary: the first one works when labels are composed of many words, the second one maps concepts with short labels (one or two words).

3.2.1 Exploiting the structure of T_{Target}

This first technique, STR_T, works on MC, the set of mapping candidates of a concept c_S. MC includes concepts with a high similarity value with c_S (only the three most similar concepts b_1, b_2, b_3 are retained) and Inc, the set of concepts of T_{Target} with a label included in the label of c_S. The idea is to exploit the location of the mapping candidates in T_{Target}. If a great number of elements in MC has a common ancestor which is deep enough in T_{Target}, that means that those elements are close and share a a common context, and we assume that c_S is meaningful according to that context too. That way we avoid mappings with isolated candidates meaningful in another context, which similarity measure is a little higher. The concept retained for the mapping with c_S belongs to the common context and has the highest similarity value. It is a possible parent or a sibling of c_S depending on whether it belongs to Inc or not.

3.2.2 Exploiting the structure of WordNet

The second technique, STR_W, exploits the hyperonymy /hyponymy WordNet structure in order to map concepts which are semantically similar without being synonyms and which labels are syntactically different. This technique can, for example, map cantaloupe with watermelon which are not synonyms but two specializations of melon in WordNet.
The use of WordNet is as follows. An expert identifies the application root node, noted $root_A$, that is the most specialized concept in WordNet which generalizes all the concepts of the concerned application domain. Then we search WordNet for the hypernyms of each term of T_{Source} not yet mapped and of each term of T_{Target} (according to all their senses) until $root_A$ or the top of WordNet is reached. Only the paths from the invoked terms to $root_A$ will be selected because they represent the only accurate senses for the application. That way, a sub-tree, called T_W, is obtained. It is composed of all the terms and the relations of the retained paths. For each concept c_S, we select in T_W the most similar concept belonging to T_{Target} using Wu and Palmer's measure [4]. This selection is very efficient because it doesn't require the computation of many similarity measures [2].

4. EXPERIMENTS AND DISCUSSION

Two kinds of experiments have been performed. First, experiments have been made in the setting of the e.dot project[1], on two real-world taxonomies in the field of predictive microbiology. Second, we applied our techniques on test taxonomies [5]. The latter are not structurally dissymmetric and cover a large domain. The application conditions of the techniques are not achieved but our objective is to make these tests in order to sketch some ideas to do improvements and to widen the scope of our approach. These experiments have shown where our specific strengths and weaknesses are. Whatever taxonomy we aligned, our approach was able to retrieve almost all the expected equivalence mappings. Furthermore, its strong point is to propose as a bonus a lot of other mappings (subclass mappings). Some mappings have a high precision and are then certain (likely mappings generated by the terminological techniques). Other ones (potential mappings generated by the structural techniques) are less certain (low precision) and have to be validated. This confirms the order in the application of our techniques. Concerning the structural techniques, STR_T proved to be very useful and leads to relevant mappings when concepts have labels composed of a lot of words and when some words are common to many labels. On the opposite, STR_W is all the more appropriate since the application domain is small. The real-world taxonomies which have motivated our approach gather all these characteristics, unlike the others. Better results are then obtained.

5. CONCLUSION

We described two structural techniques to align structurally asymmetric taxonomies. These techniques are original because different from a search of structural similarity in models. They are executed to suggest additional mappings. These mappings are not certain but they can be a good complement, if human involvement is possible, as experiments showed. We will continue this work by adapting and extending our techniques according to the experiment results. Our first objective is to be able to align taxonomies relative to larger application domains.

6. REFERENCES

[1] D. Lin. An Information-Theoretic Definition of Similarity, In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pp. 296-304, 1998.

[2] C. Reynaud, B. Safar. Structural Techniques for Alignment of Taxonomies: experiments and evaluation, In *TR 1453, LRI, Univ. of Paris-Sud*, June 2006.

[3] P. Shvaiko, J. Euzenat. A Survey of Schema-based Matching Approaches, In *Journal on Data Semantics*, 2005.

[4] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection, In *Proc. of 32nd Meeting of the Ass. for Computational Linguistics*, 1994.

[5] http://www.ontologymatching/evaluation.html

[1]E.dot is a research project funded by national network on software technology (RNTL), 2003-2005.