Structural Techniques for Alignment of Taxonomies: Experiments and Evaluation

Rapport de recherche LRI n°1453

Chantal Reynaud[1] and Brigitte Safar[1]

[1] LRI-PCRI, Bâtiment 490, CNRS & Université Paris-Sud XI - INRIA Futurs
91405 Orsay cedex
{chantal.reynaud, brigitte.safar}@lri.fr
http://www.lri.fr/~cr

Abstract. This paper deals with taxonomy alignment and presents 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. Experimental results are given and analyzed to demonstrate the effectiveness of the alignment approach. We also give comments about test cases and sketch some ideas to make improvements in order to widen the scope of the approach.

Key words. Taxonomy, Alignment, Mapping, Heterogeneous Sources, Unified Access

Résumé. Ce papier porte sur l'alignement de taxonomies et présente des techniques structurelles d'une méthode d'alignement pouvant être mises en œuvre lorsque les structures des taxonomies sont hétérogènes et dissymétriques. L'objectif est d'unifier l'accès aux documents d'un même domaine d'application, supposé s'appuyer sur des taxonomies. Nous présentons et analysons ensuite les résultats de trois expérimentations effectuées sur diverses taxonomies à l'aide du prototype développé, TaxoMap : des taxonomies réelles qui ont motivé notre approche ainsi que des taxonomies tests mises à disposition des chercheurs de la communauté.

Mots clés. Taxonomie, Alignement, Mapping, Sources hétérogènes, Accès unifié

1 Introduction

Retrieval of relevant documents is often a non trivial task. Semantic-based developments should ease this process by allowing semantic matchings between user queries and annotated documents. Moreover the number of relevant accessible documents should increase if mappings between the terms used in ontologies supporting the access to the documents were defined. Finding adequate alignment techniques on metadata or on ontology schemas is then required and has to play a major role in the next years.

Our work focuses on alignment techniques. The objective is to provide a uniform access to documents within an application domain. We assume retrieval of documents is based on very simple ontologies reduced to classification structures, i.e. taxonomies. Indeed, taxonomies are often used as a common and effective way to achieve some semantic agreements among stakeholders within a domain. The description of the content of most of today's information systems is often not specified very much. In that case, the proposed approaches which rely on OWL data representations exploiting all the ontology language features unfortunately don't apply.

Taxonomies while describing a same domain may be represented in different vocabularies and structures. They are not uniform representations. Each designer can use its own vocabulary. The model built represents his own view on the domain. Moreover knowledge is captured in an arbitrary taxonomy encoding. Encoded details may be different. Some taxonomies may be simplified or general views in comparison with others which may be more specialized. This can influence the number of levels in the representation and the size of the taxonomies.

This paper addresses the problem of alignment when the structures of the taxonomies are heterogeneous and dissymmetric, one taxonomy being deep whereas the other is flat. Such a situation can be encountered for example when we try to access to 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. The approach can also be well-suited to relate terms extracted from documents to terms in an ontology, the aim being to acquire the relevant terms in the ontology usable to semantically annotate the document before its storage in a data warehouse.

We propose alignment techniques to find mappings between taxonomies belonging to a general methodology usable across application areas. We classify the found mappings into two groups: probable mappings and mappings to be confirmed or denied

manually. The mapping process can be viewed as an execution of various techniques, invoked in sequence: terminological and then structural and semantic ones. Terminological techniques are applied first. They are principally based on the comparison of strings. They provide mappings exploiting the whole richness of the labels of the concepts. These techniques are efficient in the sense that they provide high-quality alignments corresponding to probable mappings. Unfortunately, they are not sufficient because a lot of mappings are not discovered. Our aim is to extend the handling of labels to increase overall effectiveness. Structure and dictionaries may provide additional evidence in cases where labels are not sufficient even if, here, common heuristic rules taking structure into account cannot be applied. Due to the structural dissymmetry in the mapped taxonomies, similarity of two entities cannot be identified based on the status of their respective parents and siblings. Consequently, we propose particular techniques suited to our work context, deriving interesting mappings. These mappings are identified using heuristic rules which don't provide enough support to unambiguously identify a mapping. A user evaluation is therefore necessary whereas the evaluation of mappings of the first group is not or can be done very quickly.

The paper is organized as follows. Section 2 describes the alignment approach. In section 3, we present the three structural techniques, complementary to the terminological techniques, for selecting a possible relevant candidate mapping. We evaluate the effectiveness of our algorithms on real-world taxonomies and on test taxonomies extracted from a repository about ontology matching [18]. Experimental results are given and analyzed in section 4 to demonstrate the effectiveness of the alignment approach. We also give comments about test cases and sketch some ideas to make improvements in order to be able to widen the scope of the approach. Section 5 reviews related work where section 6 concludes the paper and identifies future work.

2. The alignment approach

The objective of our approach is to generate mappings between taxonomies. Taxonomy alignment is particular because only restrictive features are usable. We cannot rely on the more complex ontology features. 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 relationship establishes links with other concepts. It is the single semantic association used in the classification. A taxonomy is generally represented by an acyclic graph where concepts are represented by nodes connected by directed edges corresponding to subclass links.

Given two structurally dissymmetric taxonomies, the objective is to map the concepts of the less structured one, called the source taxonomy T_{Source}, with concepts of the more structured one, called the target taxonomy T_{Target}. It is an oriented process from T_{Source} to T_{Target}. The alignment process aims at finding one-to-one mappings between single concepts and at establishing two kinds of relations: equivalence and subclass relations. So, we define mappings as: "Given two taxonomies, T_{Source} and

T_{Target}, mapping T_{Source} with T_{Target} means that for each concept (node) c_S in T_{Source}, we try to find a corresponding concept (node), c_T in T_{Target}, linked to c_S with an equivalence or a subclass relation.

2.1 Two types of relationships

Equivalence relationships. An equivalence relationship *is-equivalent* is a link between a concept in T_{Source} and a concept in T_{Target} with labels assumed to be similar.

Subclass relationships. Subclass relationships are usual *is-a* class links. When a concept of T_{Source} is linked to a concept of T_{Target} with such a relationship, the degree of generality of the link is assumed to be the same as the subclass link between this super-concept and other ones in T_{Target}.

2.2 A combination of techniques

Alignment is based on a similarity measure, the Lin similarity measure [8], computed between each concept c_S in T_{Source} and all the concepts in T_{Target}. This measure compares strings and has been adapted to take into account the importance of words inside expressions. Various techniques are used: terminological, structural and semantic ones (cf. Fig.1). They are applied in sequence to make the overall alignment process the most efficient as possible [7]. For each technique, the objective is to select the best concept in T_{Target} among a lot of candidates (with a similarity measure not null). This best concept is not necessarily the concept with the highest similarity measure.

TaxoMap (T_{Source}, T_{Target})
1. Compute $Sim_{LinLike}$ between each concept of T_{Source} and each concept of T_{Target}
2. For each $c_S \in T_{Source}$
3. Compute MC = *MappingCandidates*(c_S)
4. If *ProbableMapping*(c_S, MC) then stop
5. Else *Structural&SemanticTechniques*(c_S, MC)

ProbableMapping(c_S, MC)
1. Compute $C_{max} \in$ MC
2. If *TestEquivalent*(c_S, C_{max}) then return c_S *is-equivalent* C_{max}
3. Else If *NameInclusion* (c_S, C_{max}) then return c_S *is-a* C_{max}
4. Else If *SignificantlyHighest*(C_{max}, MC) then return c_S *is-a* Father(C_{max})

Fig.1. The alignment process

Terminological techniques are executed first. Being based on the richness of the labels of the concepts, they provide the most relevant mappings. Equivalence relationships are first discovered. They mapped concepts with a similarity measure corresponding to a strong similarity (greater than a threshold which has been set to 1 in our experiments). Then, we consider the inclusion between name strings. We propose a subclass mapping between c_S and c_T if c_T is the concept in T_{Target} with the highest similarity measure and if the name string of c_T is included into the name string of c_S. Finally, if the name string of the concept c_T of T_{Target} with the highest similarity measure is not included into the name string of c_S, but if its similarity measure is

significantly highest than the other ones, c_T is supposed to be a brother of c_S and the system proposes a subclass relationship between c_S and the father node of c_T. All these techniques only rely on the values of the similarity measures. They lead to mappings which are generally reliable but not always in a sufficient number.

Other techniques are needed to provide additional mappings when string comparison is not sufficient. In this paper, we focus on structural and semantic techniques. These techniques lead to identify a mapping as a correspondence between close concepts, assuming that if the suggested mapping is wrong, the right mapping establishes a relationship with another concept located in proximity in the target taxonomy. It is a guide for the user who will not have to browse the whole target taxonomy when studying the results of the system.

3. Exploiting structural features

The three techniques presented in this section take advantage of the structure of various representations. The richest knowledge representation structure is supposed to be the structure of T_{Target}. Therefore the first structural technique is performed on T_{Target}. In a second step, we propose to use an external resource, WordNet [11], exploiting its structure and its semantic relationships. Finally, in a last step, we perform treatments based on the structure of both taxonomies, being aware that T_{Source} can be very flat. Discovered mappings are essentially subclass ones.

3.1 Taking advantage of structural features in the target taxonomy

This technique works on *MC*, the set of mapping candidates of a concept c_S in T_{Source} identified from similarity measures. *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. When it has not been possible to generate a probable mapping with any element of *MC*, the idea is to exploit their position in T_{Target}.

In order to analyse the sub-graph grouping the nodes of *MC*, we compute their Lowest Common Ancestor, *LCA*, the node of highest depth that is an ancestor for all the nodes of *MC*. As an example, Fig. 2 represents the sub-graph of T_{Target} grouping the elements of $MC = \{b_1 = \text{pork meat tissue}, b_2 = \text{beef connective tissue}, b_3 = \text{beef fat}\} \cup Inc = \{\text{beef}\}$ for $c_s = \text{beef adipose tissue}$. The node Fresh meat is the *LCA* for all the elements of *MC*.

In the best case, if all the elements of *MC* are nodes having the same father in T_{Target}, the *LCA* is this father and the c_S involved concept is likely a child too. In the opposite, if all the elements of *MC* are very distant in the taxonomy, the *LCA* is a node at a high level in the taxonomy and is not an accurate father of c_S because too general. Usually, we compute partial ancestor nodes, which are *LCA* only for a subset of elements of *MC*. For each possible partial ancestor *Anc*, we compute the rela-

tive density *DR(Anc)* of the elements of *MC* having this same common partial ancestor node, MC_{Anc}. The intuition of the relative density is to take into account three criteria: (1) the number of elements in MC_{Anc}, (2) Sim_{Lin_Like}, the similarity between the elements in MC_{Anc} and c_S (3) the distance as the number of edges on the paths from each element of MC_{Anc} to *Anc*.

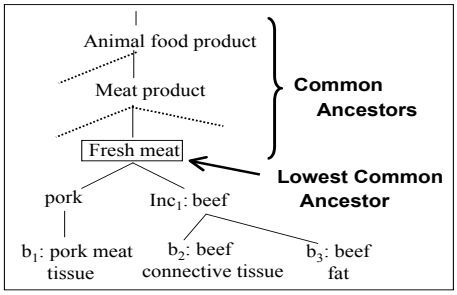

Fig.2. Sub-graph representing the elements of MC in T_{Target}

The partial ancestor with the highest relative density is the most relevant one. The relative density for a partial ancestor *Anc* is all higher since:

- *Anc* is the ancestor of a great number of elements of *MC*,
- *Anc* is the ancestor of very similar nodes to c_S ,
- Its distance to its descendants in MC_{Anc} is low.

$$DR(Anc) = \frac{|MC_{Anc}| * \sum_{Ct \in MC_{Anc}} Sim_{LIN\text{-}Like}(c_s, c_t)}{|MC| * \sum_{Ct \in MC_{Anc}} dist(c_t, Anc)}$$

In Fig.2, Fresh meat is the *LCA* of the four mapping candidates, with a distance of 7 (dist(b_1, Fresh meat) + dist(Inc_1, Fresh meat) + dist(b_2, Fresh meat) + dist(b_3, Fresh meat) that is 2+1+2+2). However, beef is the partial ancestor of three mapping candidates {beef, beef connective tissue, beef fat} with a distance of only 2 (dist(Inc_1, beef) = 0). Results given by the *DR* formula are presented in Fig.3. The relative density of beef, *DR*(beef), is the highest. Therefore, the most relevant ancestor is beef.

The most relevant ancestor can be viewed as a concept defining the context common to all its children. It is the context of a great number of mapping candidates which shares a close sense. We assume that c_S is meaningful according to that context too, avoiding mappings with concepts with a little higher similarity measure but meaningful in another context (as pork meat tissue in Fig.3).

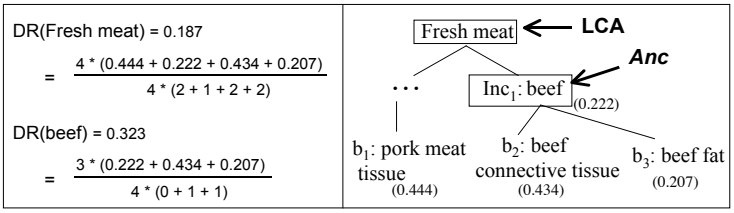

Fig.3. Results of the relative density for the elements Fresh meat and beef

Once the most relevant ancestor *Anc* has been identified, we note C_{MaxAnc} the node in MC_{Anc} with the highest similarity value. It is assumed to be the mapping candidate the most similar to c_S. If C_{MaxAnc} belongs to *Inc*, the set of concepts with a label included into the label of c_S, it is suggested as a possible father of the involved concept c_S. Otherwise, C_{MaxAnc} is proposed as a possible brother and its father (not necessarily *Anc*) will be suggested as a possible father of c_S. On Fig.3, beef is assumed to be the most relevant ancestor. beef connective tissue is the node in MC_{Anc} with the highest similarity value to c_S (c_S = beef adipose tissue*)*. That way, beef adipose tissue will be a brother of beef connective tissue and linked to beef with a subclass relationship.

3.2 Exploiting the hierarchical structure of additional background knowledge

Prior techniques are not enough if concepts are semantically the same but their label are syntactically different. For example, there is no technique to match cantaloupe with watermelon whereas querying a linguistic resource such as WordNet, an online lexical database for English language, can inform that the two concepts are a kind of melon, and then semantically very similar. In our approach, the use of synonyms is not enough. We propose to exploit the hyperonymy/hyponymy WordNet structure in order to find, for each concept of T_{Source} not yet mapped, the concept of T_{Target} semantically similar (with common ancestor nodes). This approach can map cantaloupe with watermelon which are not synonyms but two specializations of melon.

The use of WordNet is as follows. An expert identifies the application root node, denoted $root_A$, that is the most specialized concept in WordNet which generalizes all the concepts of the concerned application domain (food in the example). Then we search for the hypernyms in WordNet 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. For example, the result of a search on cantaloupe is two sets of hypernyms corresponding to two different senses.

Sense 1: cantaloupe → sweet melon → melon → gourd → plant → organism → Living thing

Sense 2 : cantaloupe → sweet melon → melon → edible fruit → green goods → food

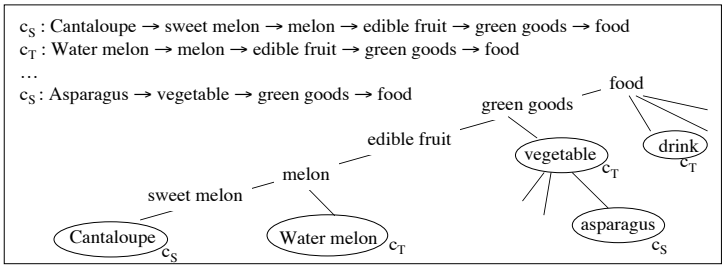

Fig.4. A sub-graph of S_{WN} where cantaloupe and watermelon are related

The paths from the invoked terms to $root_A$ will only be selected because they represent the only senses (sense 2 in the example) which are accurate for the application. That way, a sub-tree, denoted S_{WN}, is obtained. It is composed of the union of all the

terms and the relations of the retained paths (cf.Fig.4), leaf nodes coming from the two initial taxonomies (circles in Fig.4) and middle nodes being extracted from WordNet but possibly belonging to one of the taxonomies too.

For each concept c_S in T_{Source} not yet mapped, our objective is to select in S_{WN} the most similar concept belonging to T_{Target}.

Our strategy is based on Wu and Palmer's similarity measure [15]. Given two nodes c_1 and c_2, this measure gives a score depending on their depth, $depth(c_i)$, i.e. the number of edges on the path from the root of the tree to the node c_i, and also on the depth of their Lowest Common Ancestor, $LCA(c_1, c_2)$:

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * depth(LCA(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

It is more precise than a measure only based on a distance of one node to another. Indeed, the more important the depth of the *LCA* of two concepts is, greater the number of common characteristics is and the more similar concepts are. In Fig.5, if we search for the most similar concept to c_S among the nodes in S_{WN} belonging to T_{Target}, X_1, X_2, Y and Z, the similarity measures computed are in decreasing: $sim_{W\&P}(c_S, X_1)$, $sim_{W\&P}(c_S, Y)$, $sim_{W\&P}(c_S, X_2)$, $sim_{W\&P}(c_S, Z)$.

On the other hand, our strategy allows to minimize the number of similarity measures having to be computed. We elaborated our strategy from an analysis of the values of the similarity measures $Sim_{W\&P}$. According to this measure, the concept that is the most similar to a node c_S is its father, $father(c_S)$. Moreover, the similarity is higher between c_S and any of its brothers or any of the descendants close to its brothers than between c_S and its grandfather, $GF(c_S)$, until a depth p that can be computed given any node c_S in function of its depth in the tree. Indeed, in any tree, the *LCA* of a given node and of any of its ancestors is that ancestor, i.e., $LCA(c_S, father(c_S)) = father(c_S)$, $LCA(c_S, GF(c_S)) = GF(c_S)$ and the *LCA* of a given node and of any of its brothers or of any descendant of its father is that father, $LCA(c_S, desc(father(c_S))) = father(c_S)$. If n is the depth of c_S in S_{WN}, $(depth(c_S) = n)$, the depth of the *LCA* of this node and of any of its ancestors will be the depth of the considered ancestor, i.e. the depth of the node c_S minus the distance in number of edges to its ancestor: $depth(LCA(c_S, father(c_S))) = n-1$ and $depth(LCA(c_S, GF(c_S))) = n-2$.

The similarity of c_S with its grandfather, $GF(c_S)$, is then defined by:

$$Sim_{W\&P}(c_S, GF(c_S)) = \frac{2 * depth(LCA(c_S, GF(c_S)))}{depth(c_S) + depth(GF(c_S))} = \frac{2 * (n - 2)}{n + (n - 2)} = \boxed{\frac{n - 2}{n - 1}}$$

The similarity of c_S with one of its brothers or any descendant of its father with a depth p is defined by:

$$Sim_{W\&P}(c_S, desc(father(c_S))) = \frac{2 * depth(father(c_S))}{depth(c_S) + depth(desc(father(c_S)))} = \boxed{\frac{2 * (n - 1)}{n + p}}$$

The solution of the following inequation indicates the value of p from which the similarity of the grandfather will be higher than the similarity of any descendant of the father.

If $n > 2$,

$$\frac{n-2}{n-1} > \frac{2*(n-1)}{n+p} \Leftrightarrow (n-2)(n+p) > 2(n-1)^2 \Leftrightarrow (n+p) > \frac{2(n-1)^2}{n-2}$$

$$\Leftrightarrow p > \frac{2(n-1)^2}{n-2} - n \Leftrightarrow p > \frac{2n^2+2-4n-n2+2n}{n-2} \Leftrightarrow \boxed{p > \frac{(n-1)^2+1}{n-2}}$$

If $n = 2$, $Sim_{WP}(c_S, GF(c_S)) = 0)$ et $\forall p$, $Sim_{WP}(c_S, GF(c_S)) < Sim_{WP}(c_S, desc(father(c_S)))$

In the same way, we can compute the depth p' from which the similarity of the great-grandfather must be considered, and so on.

Once these limits have been computed, the search strategy of the term of T_{Target} the most similar to a given element c_S in T_{Source} is the following. At first, we test if the father of c_S in S_{WN} belongs to T_{Target}. If it is, the father of c_S is the most similar term to c_S according to Wu&Palmer's measure. Otherwise, we search for a node belonging to T_{Target} being a descendant of the father of c_S, and having a depth lower than p (or equal). If no node holds, we test the grandfather, then in an alternative way the descendants of the father with a depth $p+1$ and the direct-descendants of the grandfather. The alternative test process is reiterated by increasing the depth of the tested nodes every iteration until depth p' is reached. Then we test the great-grandfather and again new alternative tests are done in each direction. As soon as a concept belonging to T_{Target} is found, we have to verify if other elements belonging to T_{Target} are direct-descendants of the last explored nodes in the other directions. If there are some, we compute their similarity measure with c_S and we retain the node with the highest measure. Otherwise, the unique term belonging to T_{Target} is retained. We can remark that very few similarity measures must be computed.

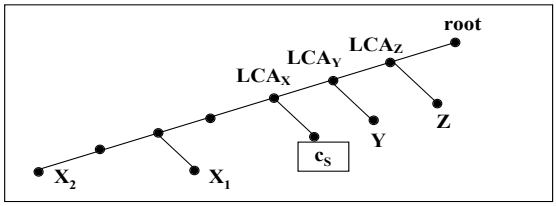

Fig.5. Example of a S_{WN} tree

On Fig.5, the father of c_S does not belong to T_{Target}. The depth of c_S is 4, the depth of the grandfather is 2, the limit depth p for c_S is 5. No descendant of the father of c_S with a depth lower than 5 belongs to T_{Target}. Neither does the grand-father of c_S. p' associated to the great-grandfather with a depth of 1 is 11. Descendants of the father of c_S with a depth of 6 are then tested. X_1 belongs to T_{Target}. Then we search for other possible candidates, direct nodes of already explored nodes with a depth lower (or equal) than p'. Y belongs to T_{Target}, however $sim_{W\&P}(c_S, X_1) = 0,6$ while $sim_{W\&P}(c_S, Y) = 0,57$. Consequently X_1 is the most similar concept to c_S.

This technique allows to establish mappings between concepts known in WordNet, i.e. labelled by expressions generally composed of only a few words. No precision on the kind of relationship between the two concepts can be given

3.3 Exploiting structural features in both taxonomies

At this point, we propose to apply heuristics similar as those proposed in [9], [10], [12]. The basis idea is to make suggestions based on the mappings of adjacent nodes prior established. In the example described Fig. 6, the problem is to find a mapping for Apple Cider with 12-14 Brix, a child of Fruit and fruit products in T_{Source}. As a great part of the children of Fruit and fruit products in T_{Source} have been mapped with Drink or with a more specialized node in T_{Target}, Apple Cider with 12-14 Brix may be mapped with a node of the sub-tree rooted in Drink. So, the problem is to identify a general node in T_{Target} similar to c_S, then if c_S must be mapped with that general node (Drink in Fig. 6) or with a more specialized one (for example Apple juice in Fig. 6).

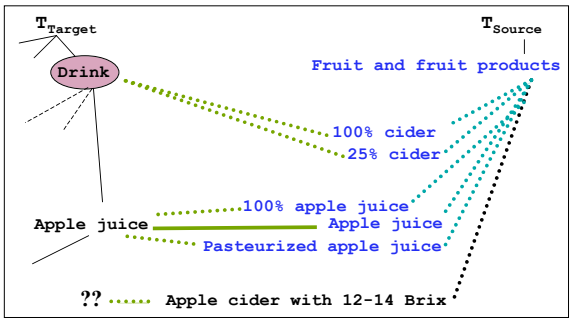

Fig. 6. Mappings of brothers of Apple cider with 12-24 Brix

Given c_S a leaf in T_{Source}, we define MappingsOfNeighbours (MoN) as the set of nodes in T_{Target} composed of the nodes mapped with the brothers of c_S. For each element of MoN, we compute the number of mappings established with a brother of c_S. Only elements of MoN which are mapped at least twice are retained. They are elements of CMoN. Fathers of elements of CMoN are relevant general nodes only if the number of their children in CMoN is more than 1/3. Elements of MoN are presented to the expert for validation grouped by general nodes. The groups are ordered in decreasing according to the number of established mappings.

In Fig.7, 3 among the 11 brothers of reduced-fat containing egg yolk have been mapped with egg, 4 with egg based product, 2 with sauce, 2 with mayonnaise. In T_{Source}, egg and egg based product have a common father Egg and Egg products. The father of sauce and mayonnaise is Mustard, condiments, spices. The system elaborates two propositions. Each proposition is a suggestion for a mapping either with the general node (i.e. Egg and Egg product or Mustard, condiment, spices*)* or with one of its specializations, without giving any information on the kind of relationships between the two mapped elements.

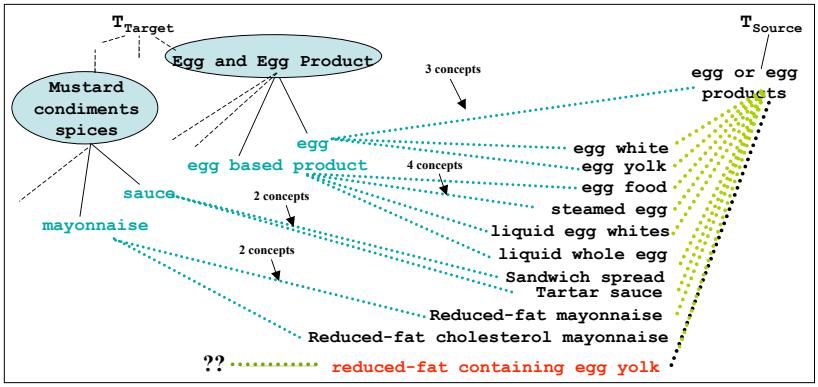

Fig. 7. Mappings of the brother nodes of reduced-fat containing egg yolk.

If c_S is not a leaf in T_{Source}, the search of the elements of MoN is done on the children of c_S, its brothers, their descendants. A mapping will be proposed with the element of CMoN which is the lowest common ancestor of CMoN.

This technique gives relevant results when mappings have been prior generated for a lot of brother or children nodes, even if the taxonomies to match are very different from a structure point of view.

4. Experimentations

The mapping techniques described above are regarded as independent components that made up Taxomap, a prototype implemented in java. Two kinds of experiments have been performed. First, experiments have been made on two real-world taxonomies, Sym'Previus and Com'base, in the field of predictive microbiology in the setting of the e.dot project[1]. Second, we applied the techniques on test taxonomies extracted from a repository about ontology matching [18]. These last taxonomies are not structurally dissymmetric and cover a larger domain. The application conditions of the techniques are not achieved but our objective is to test them in order to sketch some ideas to do improvements and to widen the scope of our approach.

4.1. Experiments in the field of predictive microbiology

The objective in the e.dot project was to access to Sym'Previus and Com'Base information sources, both containing documents in the field of predictive microbiology, using the querying system MIEL. MIEL is only based on Sym'Previus concepts organized in a hierarchically well-structured taxonomy. Accordingly, the access can be

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

achieved only if mappings between Com'Base and Sym'Previus concepts are defined. In this alignment process, Sym'Previus is T_{Target}, Com'Base is T_{Source}.

Sym'Previus taxonomy is structured in a hierarchy composed of around 400 concepts related by subclass relationships. It has seven layers. Com'Base consists of 172 concepts related by subclass links. The concepts are organized in a hierarchy which is not very structured: only two layers depth, the first level being composed of only 12 concepts. The two hierarchies have overlapping parts but they are structured in a different way. In one hand, they represent different standpoints. On the other hand, they have been designed independently and they have arbitrary granularity.

Sym'Previus and Com'Base taxonomies are relative to a very specific, restricted and refined application domain. Labels of concepts are rather long, composed of several words. Often, some labels include other ones, words being added to the label of a concept to obtain the label of a more specialized one. So, identical words may be present in a lot of labels. That way, a lot a similarity measures are not null because labels share common words or strings. The problem is to select the most relevant concept among all the mapping candidates (concepts in T_{Target} having a similarity measure different from zero) of each concept c_S in T_{Source}. For evaluation purpose, we asked the knowledge engineer to manually establish the mappings between the two taxonomies, i.e. for 172 conceps (the number of concepts in Com'Base). 44 equivalence mappings and 121 subclass mappings have been proposed by the expert (a subclass mapping establishes a subclass relationship between a concept c_S of T_{Source} and a concept c_T in T_{Target}, c_S being a brother different from the other children of c_T). No mapping has been proposed for 7 concepts (middle nodes in T_{Source}).

96 relevant mappings have been generated by our system among 101 probable proposed ones. These mappings have a high precision, greater than 90% [7] to the detriment of recall (58 %). Structural techniques are then very useful to complete these results, even if the obtained mappings are less sure. Their precision, according to our experiment results, confirms the order in the application of the techniques. Obtained results are summarized in Table 1.

Structural and semantic techniques are applied to the 64 concepts not yet mapped. For the technique exploiting the structure of T_{Target}, problems are encountered with strings composed of many words. Indeed, this technique exploits the structure of T_{Target}, but it is greatly based on the similarity measures between concepts, these measures being principally based on string comparisons. The similarity measure applied on long strings with only one word (sometimes very small) referring to the underlying concept and a major part composed of words which precise and characterize this underlying concept and which belongs to labels of concepts in T_{Target} leads to bad results. The similarity measure may also be irrelevant when few words of the label of a concept in T_{Source} are words of labels of concepts in T_{Target}. These problems excluded, the technique exploiting the structure of T_{Target} was very useful in 28 cases among 42, i.e. in 66 % of the cases.

Table 1. Number of obtained mappings per structural technique

Structural Techniques	# studied terms	# proposed mappings	# confirmed mappings	# non found mappings	Precision
Exploiting T_{Target} structure	64	42	28	22	66 %
Exploiting WordNet	22	15	9	7	60 %
Exploiting T_{Target} & T_{Source} structures	7	5	3	2	60 %

The technique based on WordNet was an interesting complement of the techniques prior applied. 15 mappings have been generated. For example, 100% cider was mapped to Drink and Frankfurter to Sausage. The 7 concepts with no suggestion are acronyms (example: TSB), technical or too long concepts non recognized by WordNet (example: Egyptian Kofta). Among the false suggestions, for example, Lamb is mapped to Meat whereas the expert proposed to map it with to Sheep, a more specific concept.

We applied the last technique on the 7 last concepts no yet mapped. Two concepts, TSB and Phosphate buffer, have not been mapped because they have not enough brothers to make this technique applicable. For the five other concepts, 3 relevant propositions are made. The suggestions are to map Tampeh, Brocoli, Pecan and Pecan nuts with Fresh Fruits and Vegetables, (for Pecan and Pecan nuts the expert proposed a mapping with an other children of Vegetable, Dried Fruits and Vegetables). For the last concept, Egyptian Kofta, the system proposes 3 matching directions, one with Fresh meat, another with Meat-based product (which is the proposition of the expert) and the third one with Poultry.

4.2. Experiments on test taxonomies

Test taxonomies extracted from a repository about ontology matching [18] have been run. Experiment results demonstrate the effectiveness of our approach and give ideas to make improvements in order to be able to cover other kinds of taxonomies. Indeed, the characteristics of these test taxonomies are different from the characteristics of Sym'Previus and Com'Base having motivated our approach.

Experiments on Russia taxonomies: These two taxonomies, Russia-A and Russia-B, cover a very large domain describing Russia, its geography and its monuments [17]. They have approximately the same number of concepts (300). They have the same depth (7 layers depth). Furthermore labels are very often composed of a single word. All these features make them very different from Sym'Previus and Com'Base and influence the alignment process in a different way.

96 equivalence mappings among 103 expected mappings have been generated by TaxoMap. The 7 concepts with no mappings should have been mapped with concepts semantically equivalent having a different label. Furthermore, TaxoMap proposed 29 relevant additional subclass mappings. These results are satisfactory because a lot of expected mappings are retrieved. Yet, these results don't bring to the fore the strengths of our approach. When labels are composed of a unique word, this word generally differs from one label to another. Consequently, labels of concepts linked by a subclass relationship rarely share common words. During the alignment process, a concept c_S in the source taxonomy has a very limited number of mapping candidates in T_{Target}. The approach whose aim is to select the most relevant concept among a lot of mapping candidates, assuming that the number of elements in the set of candidates is at least 3, is then very often inoperative.

Under these conditions, good results are obtained when mapping candidates are at least 3 (rare). For example, the technique based on the inclusion between name strings

when the included concept has the highest similarity measure allows identifying about 15 relevant mappings, such as Azov_sea or black_sea is-a sea, capital_city is-a city, cathedral_of_sophia is-a cathedral, or monetary_unit is-a unit. These mappings are not expected mappings provided with the test taxonomies which are only equivalence mappings but they are all the same relevant. On the other hand, when the included concept is the unique mapping candidate, irrelevant mappings are often generated, as North_America is-a North or Easter is-a East.

The technique based on WordNet is inadequate to align these taxonomies because of the coverage of the domain which is too large. It builds a sub-tree from all the hypernym nodes in WordNet until the most general concept in the application is reached. When the domain is very large, the most general concept is the root node in WordNet. Consequently, S_{WN} is very big. It mixes up various senses and leads to irrelevant suggestions. Improvements could be obtained if several sub-trees are built, one per sub-domain assuming that the various sub-domains can be known.

The last structural technique exploiting the mappings of the brother nodes of the involved concept in T_{Source} is neither well-suited because concept nodes in T_{Source} have very few brothers.

Improvements are possible to obtain more relevant results. Currently, when there are only one or two mapping candidates, the concept with the highest similarity value is chosen. That way, bus was considered as a brother of foreign_business_person and Chechnya was considered as a brother of Check whereas the two similarity measures were very low. This experiment proves the need to reject mappings with a too low similarity measure (less than a given threshold) or to try to confirm them with an additional technique. For example, querying WordNet could provide very useful additional semantic knowledge.

Experiments on Course Catalog taxonomies: These two taxonomies cover a very large domain too. They contain course information from two universities, Cornwell and Washington [16]. Once more, they have approximately the same number of concepts (150). They have the same depth (4 layers depth). However, unlike the previous taxonomies, a lot of labels are composed of several words. A lot of words are contained in a lot of labels, consequently the mapping candidates of an involved concept c_S in T_{Source} are more numerous. Our techniques can operate to determine the most relevant concepts.

50 expected equivalence mappings are provided. 45 have been recognized by TaxoMap. In 35 cases, the target concept is considered as equivalent, in 10 cases the target concept is the most probable brother. 77 additional mappings are proposed by TaxoMap, 52 are correctly located according to a manual validation.

Techniques providing probable mappings have a very good precision. The inclusion between terms leads to identify 9 subclass relationships between concepts of Washington and of Cornwell and all these mappings are relevant (for example: Applied Mathematics is-a Mathematics, French Linguistics is-a Linguistics). The technique based on the significantly highest similarity measure leads to establish 15 additional relevant mappings among 16 that are proposed (for example: Political_Science is considered as a

brother of Political_Theory, International_Studies_Jewish_Studies is considered as a brother of Program_of_Jewish Study). The high precision proves that these mappings are sure.

In this experiment, structural techniques have been often used to discover additional mappings. The structural technique performed on T_{Target} is applied 43 times and leads to 24 relevant mappings. For example, Ancient_and_Medieval_History is related to Medieval_Renaissance, and to Early_Modern_European_History, two sub-classes of History, Biology is close to Plant_Biology and to Microbiology, having as common parent concept Decision of Biological Science. Yet, the precision of the technique is not so good as in our first experiment for various reasons: the coverage of the domain which is larger, less refined labels, semantics of concepts not only given by their name which are not very expressive but also by their location in the taxonomy, techniques applied in sequence. As the coverage of the domain is larger, concepts are more general and have to be interpreted in the context of the hierarchy. For example, Literature and Language_course are rather general concepts but in the setting of the Course Catalog, they must be interpreted in the context of Near_Eastern_Studies. Our approach doesn't exploit simultaneously all structural information, then irrelevant mappings may be generated. For example, we could wish to map Slavic_languages_and_Literatures to Russian_Language which belongs to its mapping candidates set. Instead, TaxoMap proposes to locate this concept close to Literature and Language_Course by establishing a subclass mapping with their common parent concept Near_Eastern_Studies. Of course, this is irrelevant. Finally, as for the previous test case, the application domain is too large and the concepts in T_{Source} have too little children to make the two last structural techniques operative.

These experiments have shown us where our specific strengths and weaknesses are, and how we can continue on improving. The approach is suitable for very refined taxonomies containing only subclass relationships. On the other hand, it is less appropriate to general taxonomies modeling implicit various relationships such as part_of, is-a, instance_of, made_of, and so on. Yet, whatever taxonomy we align, our approach was able to retrieve almost all the expected equivalence mappings. Furthermore, the strong point of TaxoMap is to propose in addition a lot of additional relevant mappings (+ 29 in Russia, + 52 in Course Catalog, the same number as the expected mappings). Some of them have a high precision and are then sure (generated by the terminological techniques). Other ones (generated by structural techniques) are less sure (low precision) and must be validated but if human involvement is possible, the approach is very interesting because much more mappings can be obtained.

Table 2. Number of obtained mappings generated from the test taxonomies

	Course Catalog		Russia	
	T_{Target}	T_{Source}	T_{Target}	T_{Source}
	Cornwell	Washington	Russia-A	Russia-B
Number of concepts	176	167	372	310
Number of expected mappings	50		103	
Number of relevant found mappings	45		96	
Number of additional relevant mappings	52		29	

5. Related work and discussion

Currently, a lot of works aim at automating generation of mappings. A survey of these techniques is presented in Rahm and Bernstein [12] and Shvaiko and Euzenat [13]. Techniques are multiple. We only focus, in this section, on research work relative to structural techniques, central in this paper.

Structural techniques exploit the structure of compared schemas, often represented as graphs. Algorithms implementing these techniques are based on heuristics. Heuristics consider, for example, that elements of two distinct schemas are similar if their direct sub-concepts, and/or their direct super-concepts and/or their brother concepts are similar [2], [11], [1]. These structural techniques can be based on a fixed point [9]. In S-Match [4], the matching problem is viewed as a satisfiability problem of a set of propositional formula. Graphs and mappings to test are translated in propositional formula considering the position of the concepts in the graph and not only their label.

Our work is different from these ones in particular because of the dissymmetry in the structure of the taxonomies. We can't search similar structures. So, we propose to exploit structural data in a different way. The structure of the target taxonomy, solely considered, is used to determine the most relevant concept able to be mapped with a concept c_S of T_{Source}. This technique exploits the structure of the target taxonomy but it is also based on the similarity measures prior built, based on string comparisons.

As the structure of the source taxonomy is supposed to be little structured in our setting, another solution is to consider the structure of additional resources different from the matched taxonomies, for example WordNet. The use of WordNet in alignment research work is not new. A lot of alignment systems use external linguistic resources. Yet, our approach exploits WordNet in a non-common way. WordNet is not considered simply as a source of synonyms, hypernyms or hyponyms. It provides a structural support exploited to detect relations between concepts. This can be compared to what is done in CMS [5], a structure matching system implementing a series of mapping techniques. WordNet is used in CMS by the *WNNameMatcher* which exploits also the WordNet hierarchy in order to compute a similarity measure between concept pairs. The difference between TaxoMap and CMS on this point is that, in our approach, we build a unique WordNet sub-tree S_{WN} from the concepts in T_{Source} not yet mapped and from all the concepts in T_{Target}. To build S_{WN}, we only have to compute $|T_{Source}| \cup |T_{Target}|$ union operations made on the sets of retained hypernyms. Moreover, we showed in section 3.2 that using the sub-tree S_{WN} and the Wu and Palmer's measure, the search for the concept c_T belonging to both S_{WN} and T_{Target} the most similar of a concept c_s doesn't need to compute a lot of measures. On the other hand, CMS arranges each pair of compared elements in the WordNet hierarchical structure and then measures the similarity between them. So, it computes $|T_{Source}| \times |T_{Target}|$ measures.

Finally, in our approach, we propose a last technique based on the structure of both taxonomies, being aware of a dissymmetry in their structure. The idea is to rely on prior mappings to deduce additional suggestions of mappings. The technique takes

into account the location of the concepts prior mapped in each of the taxonomies and gives a great importance to the neighbourhood of the concepts. This notion of neighbourhood has been considered in other research works. "Two nodes match if nodes in their neighbourhood also match" is a widely used constraint where the neighbourhood is defined to be the children, the parents, or both [8], [9], [11]. We propose to use heuristics close to this constraint.

6. Conclusion

This paper describes three structural techniques to align taxonomies supposed to be asymmetric from a structure point of view. In our setting, we can't search identical structures. So, we propose other ways to exploit this kind of information: exploitation of the structure of the target taxonomy solely, exploitation of the structure of the hyperonymy/hyponymy hierarchy in WordNet, exploitation of the structure of both taxonomies combined with the exploitation of prior identified mappings. These techniques are original because they distinguish from a search of structural similarity in models. They are applicable to suggest mappings. These mappings are not so sure than mappings generated by terminological techniques, this explains why terminological techniques are proposed to be used first. Nevertheless, it is a good complement as experiments show it.

We will continue this work by building a toolbox proposing our techniques and other ones, each one being suitable to particular taxonomies according to their features. Our techniques are well-suited to align structurally dissymmetric taxonomies covering a restricted and refined domain and composed of concepts with labels which are expressions of several words. Other techniques are needed to widen the scope of our approach. They may be adapted from our original techniques or extended in order to take into account the features of the taxonomies being aligned: the coverage of the application domain, the complexity in the expressions to name the concepts, the depth levels, the number of equivalent terms, the user involvement and more generally all characteristics automatically manageable. The needed adaptations are of two kinds. It can be adaptations of the techniques themselves or of their use. Indeed, they can be either performed in sequence or combined.

6. Acknowledgments

We wish to thank Hassen Kefi and Ahlem Slimi for their contribution to this work.

References

1. Bach, T.-L., Dieng-Kuntz, R., Gandon, F.: On Ontology Matching Problems for building a Corporate Semantic Web in a Multi-Communities Organization. ICEIS (4) (2004), 236-243
2. Do, H. H., Rahm, E.: COMA – A system for flexible combination of schema matching approaches. VLDB. (2001) 610-621
3. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Learning to map between ontologies on the semantic web. WWW. N.Y. USA. ACM Press. (2002) 662-673
4. Giunchiglia, F., Shvaiko, P.: Semantic Matching. The Knowledge Engineering Review. (2004) 18(3):265-280
5. Kalfoglou, Y., Hu, B.: CROSI Mapping System (CMS) Results of the 2005 Ontology Alignment Contest. Integrating Ontologies Workshop., K-Cap Conference. Banff. Canada (2005) 77-84
6. Kéfi, H. : Ontologies et aide à l'utilisateur pour l'interrogation de sources multiples et hétérogènes. PhD Thesis. Université Paris Sud. March 2006.
7. Kéfi, H., Safar, B., Reynaud, C.: Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes. RFIA. Tours (2006)
8. Lin, D.: An Information-Theoretic Definition of Similarity. ICML. Madison. (1998) 296-304
9. Madhavan, J., Bernstein, P. A. , Rahm, E.: Generic matching with Cupid. VLDB Journal. (2001) 49-58
10. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity Flooding: A versatile Graph Matching Algorithm and its application to schema matching. ICDE. San Jose CA. (2002) 117-128
11. Miller, G. A.: WordNet: A lexical Database for English. Communications of the ACM. (1995) Vol. 38 n°11 39-45
12. Noy, N. F., Musen, M. A.: Anchor-Prompt: Using non-local context for semantic matching. Workshop on Ontologies and Information Sharing at IJCAI-2001. Seattle. WA. (2001)
13. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. VLDB Journal: Very Large Data Bases. (2001) 10(4): 334-350
14. Shvaiko, P., Euzenat, J.: A survey of Schema-based Matching Approaches. Technical Report DIT-04-087. Informatica e Telecomunicazioni, University of Trento (2004)
15. Wu, Z., Palmer, M.: Verb semantics and lexical selection. Computational Linguistics. Las cruces (1994) 133-138
16. http://anhai.cs.uiuc.edu/archive/domains/course_catalog.html
17. http://www.atl.external.lmco.com/projects/ontology/i3con.html
18. http://www.ontologymatching/evaluation.html