Techniques structurelles d'alignement pour portails Web

Chantal Reynaud, Brigitte Safar

Université Paris-Sud XI, CNRS (L.R.I.) & INRIA (Futurs)
91405 Orsay cedex
{chantal.reynaud, brigitte.safar}@lri.fr
http://www.lri.fr/~cr

Résumé. Le travail décrit dans cet article a pour objectif d'unifier l'accès aux documents d'un domaine d'application. Il permet en particulier d'augmenter le nombre de documents accessibles à partir de portails Web sans en modifier l'interface d'interrogation. L'accès aux documents est supposé s'appuyer sur des taxonomies que nous proposons d'aligner. Cet article porte spécifiquement sur les techniques structurelles mises en œuvre, qui sont originales et particulières dans la mesure où elles sont adaptées au traitement de taxonomies dont les structures sont hétérogènes et dissymétriques. Nous présentons et analysons ensuite les résultats de trois expérimentations effectuées sur diverses taxonomies, des taxonomies réelles qui ont motivé notre approche ainsi que des taxonomies tests mises à disposition des chercheurs de la communauté.

1 Introduction

La recherche de documents pertinents sur le Web est une tâche encore souvent laborieuse. Le Web sémantique devrait faciliter ce travail en réalisant un appariement sémantique entre la requête de l'utilisateur et les documents indexés. Les techniques d'alignement de méta-données ou d'ontologies sont pour cela de la plus grande importance.

Notre travail porte sur de telles techniques, utilisables dans le contexte du Web. Notre objectif est de permettre un accès unifié via le Web aux documents d'un même domaine d'application. Plus précisément, nous proposons d'aligner la taxonomie d'un portail Web avec celle de documents externes de façon à augmenter le nombre de documents accessibles à partir de ce portail sans en modifier l'interface d'interrogation. La recherche de tels documents s'appuie en général sur des ontologies très simples, souvent réduites à des hiérarchies de classification, c'est-à-dire des taxonomies. Les approches basées sur des représentations plus fines, telles des représentations OWL, exploitent toute la richesse du langage. Elles ne sont pas adaptées au traitement de simples taxonomies.

Les taxonomies à apparier, tout en représentant la structure d'un même domaine, peuvent comporter des termes différents ou des termes identiques structurés différemment. Chaque concepteur utilise son propre vocabulaire. Il ne s'agit pas de représentations uniformes mais de vues spécifiques à chaque concepteur, certaines étant des descriptions avec un niveau de détail important, des concepts très spécialisés y sont alors représentés, d'autres étant des descriptions plus macroscopiques et se limitant à la représentation de concepts généraux. Selon le cas, la profondeur et la taille des taxonomies varient.

Cet article présente des techniques d'alignement particulières dans la mesure où elles sont adaptées au traitement de taxonomies dont les structures sont hétérogènes et dissymétriques. En effet, si la taxonomie de concepts d'un portail Web est en général bien structurée, celle des autres documents auxquels on souhaiterait avoir accès via ce portail ne l'est pas toujours. Ce phénomène se rencontre aussi dans d'autres applications. Ainsi, l'approche peut être utile pour lier des termes isolés, extraits de documents, à ceux d'une ontologie, dans un processus d'annotation sémantique de documents avant stockage dans un entrepôt thématique. Les techniques proposées sont ainsi génériques, utilisables dans des contextes variés. Etant donné deux taxonomies structurellement dissymétriques, l'objectif est de mettre en correspondance les concepts de la taxonomie la moins structurée, la taxonomie source (T_{Source}), avec les concepts de la taxonomie la plus structurée, la taxonomie cible (T_{Cible}). Le processus d'alignement est donc orienté de T_{Source} vers T_{Cible}.

Les techniques décrites s'intègrent dans une approche plus globale de génération de mises en correspondance entre concepts, ou mappings, de deux sortes : des mappings probables et des mappings potentiels qu'un expert doit confirmer (Kéfi, 2006). Le processus d'alignement est semi-automatique. Il peut être vu comme une application séquentielle de différentes techniques : terminologiques puis structurelles. Les techniques terminologiques, basées principalement sur des comparaisons de chaînes de caractères, sont appliquées en priorité. Elles exploitent toute la richesse des noms des concepts. Ces techniques sont efficaces. Elles fournissent des mappings de grande qualité que nous qualifierons de probables. Même si elles sont efficaces, les techniques terminologiques ne peuvent cependant pas trouver l'ensemble des rapprochements possibles. Notre objectif est alors de compléter les techniques terminologiques par d'autres techniques basées sur l'exploitation de la structure. Les règles heuristiques communément utilisées consistant à considérer que deux entités de deux taxonomies sont similaires si leur voisinage respectif est similaire sont ici inapplicables. Nous proposons des techniques structurelles différentes, plus adaptées à la spécificité des taxonomies que nous cherchons à aligner. Les mappings supplémentaires générés sont moins sûrs que ceux générés par les techniques terminologiques, nous les qualifierons de mappings potentiels. Leur validation est indispensable. Trois techniques structurelles sont proposées, basées sur des éléments de structure différents, mais ne consistant en aucun cas à rechercher des similarités structurelles entre les deux taxonomies, ce qui en fait toute leur originalité.

Ce papier est organisé de la façon suivante. Dans la section 2, nous décrivons les travaux proches réalisés dans le domaine. La section 3 détaille l'approche d'alignement. Les techniques structurelles sont décrites en section 4. La section 5 porte sur les expérimentations réalisées. Enfin, nous concluons et présentons quelques perspectives.

2 Etat de l'art

De nombreux travaux portent aujourd'hui sur l'alignement d'ontologies. Une synthèse des techniques est présentée par Kalfoglou et Schorlemmer (2003) ainsi que par Shvaiko et Euzenat (2004). Les techniques sont variées. Elles exploitent différents types d'information, les noms des éléments, les types des données, la structure de la représentation des éléments des schémas, les caractéristiques des données, etc. (Madhavan et al. (2001), Yan et al. (2001), Do et Rahm (2001), Bach et al. (2004)). Dans cette section, nous nous limiterons aux travaux portant sur des techniques structurelles appliquées au niveau schéma.

Les techniques structurelles exploitent la structure des schémas comparés, souvent représentés sous forme de graphes. Les concepts ne sont pas étudiés séparément. Dans une taxonomie, ils sont considérés en tenant compte de leur position dans la hiérarchie des concepts. Les algorithmes implémentant ces techniques sont basés sur des heuristiques qui considèrent, par exemple, que des éléments de deux schémas sont similaires si leurs sous-concepts directs et/ou leurs super-concepts directs et/ou leurs concepts frères sont similaires. (Do et Rahm (2001), Noy et Musen (2001), Bach et al. (2004), Maedche et al. (2002)).

Ainsi, Noy et Musen dans Anchor-PROMPT rapprochent des ontologies vues comme des graphes au sein desquels les nœuds sont des classes et les liens sont des propriétés. Le système prend en entrée un ensemble d'ancres qui sont des couples d'éléments liés. Un algorithme analyse les chemins dans le sous-graphe délimité par les ancres et détermine quelles classes apparaissent fréquemment dans des positions similaires sur des chemins similaires. Ces classes correspondent ainsi vraisemblablement à des concepts similaires (Noy et Musen (2001)).

Les techniques structurelles peuvent être basées sur un point fixe. Dans le système (Madhavan et al. (2001)) implémentant l'algorithme Similarity Flooding (SF), l'objectif est de mettre en correspondance des schémas convertis en graphes orientés et étiquetés. Des éléments de deux schémas distincts sont dits similaires si leurs éléments adjacents sont similaires. La similarité entre deux éléments est propagée aux voisins respectifs (les éléments reliés par des arcs) jusqu'à l'obtention d'un point fixe. Des mappings sont générés pour tous les nœuds dont la valeur de similarité est supérieure à un certain seuil. L'algorithme exploite principalement les étiquettes des arcs, pour déterminer à chaque itération le poids et les nœuds vers lesquels il doit propager les similarités. Il ne donne pas de bons résultats lorsque l'étiquetage est inexistant ou lorsque les labels sont très souvent identiques.

Dans S-Match (Giunchiglia et Shvaiko (2003)), le problème d'appariement entre deux nœuds est vu comme un problème de satisfiabilité de formules de la logique propositionnelle. Les graphes et les mappings à tester sont traduits en formules logiques en prenant en compte la position des concepts dans le graphe et pas seulement leur label.

Notre travail de recherche se différencie des travaux cités précédemment, en particulier à cause de la dissymétrie dans la structure des taxonomies. Il est impossible de chercher à retrouver des similarités structurelles dans les deux taxonomies. Nous proposons alors des techniques différentes exploitant la structure de différentes représentations. La structure de représentation des connaissances la plus riche est supposée être celle de la taxonomie cible. La première technique appliquée s'appuie donc sur elle.

Nous proposons ensuite d'utiliser WordNet (Miller, 1995), en exploitant sa structure et ses relations sémantiques. Son utilisation n'est pas novatrice en alignement. Cependant, notre approche exploite WordNet d'une manière originale. WordNet n'est pas considéré uniquement comme un moyen de fournir des synonymes, des hypernymes ou des hyponymes. Il fournit un support structurel exploité pour détecter des relations entre concepts. La technique peut être comparée à ce qui est fait dans CMS (Kalfoglou et Hu (2005)), un système faisant de l'alignement structurel implémentant toute une série de techniques de mappings. WordNet est utilisé dans CMS par le module *WNNameMatcher* qui exploite aussi la hiérarchie WordNet pour calculer une mesure de similarité entre couples de concepts. La différence entre TaxoMap et CMS est que, dans notre approche nous construisons un unique sous-arbre WordNet, T_{WN}, alors que CMS construit une hiérarchie à partir de WordNet pour chaque couple d'éléments comparés. La hiérarchie WordNet est

ensuite utilisée différemment. Pour chaque c_S de T_{Source} non encore mis en correspondance, nous recherchons le concept c_T appartenant à la fois à T_{WN} et à la taxonomie cible le plus similaire. Nous montrons en section 4 que ce processus est peu coûteux d'un point de vue complexité de calcul.

Enfin, dans un dernier temps, nous proposons d'exploiter la structure de la taxonomie source combinée à celle de la taxonomie cible, sachant que la structure de la taxonomie source peut être très peu structurée. L'idée est de se baser sur des mappings antérieurs pour déduire des suggestions supplémentaires de mappings. Cette technique prend en compte la localisation des concepts préalablement mis en correspondance dans chacune des taxonomies et donne une grande importance au voisinage des concepts. Cette notion de voisinage a été exploitée dans d'autres travaux de recherche. « Deux nœuds sont similaires si leur voisinage est similaire » est une contrainte très largement utilisée où l'entourage est défini comme étant composé des enfants, des parents, ou de l'ensemble des deux (Madhavan et al. (2001), Melnik et al. (2002), Noy et Musen (2001)). La troisième technique structurelle mise en œuvre utilise des heuristiques proches de cette contrainte.

3 Approche d'alignement

Les schémas en entrée du processus de mise en correspondance sont des taxonomies, correspondant à des ontologies très sommaires avec des définitions de concepts très pauvres. Les concepts sont principalement définis par référence à leur terminologie. Ils n'ont pas d'attributs.

Une taxonomie (C, H_C) comprend un ensemble de concepts C et une hiérarchie de subsomption entre concepts H_C. Un concept est défini par son label et les relations de sous-classes qui le relient à d'autres concepts. Le label est un nom (chaîne de caractères) qui décrit des entités en langage naturel et qui peut être une expression composée de plusieurs mots. Les relations de sous-classes établissent des liens entre concepts. Il s'agit de l'unique association sémantique utilisée dans la classification. Une taxonomie est généralement représentée par un graphe acyclique dont les nœuds sont les concepts et les arcs correspondent aux liens de sous-classes.

3.1 Caractéristiques des taxonomies alignées

Dissymétrie : Les techniques que nous proposons sont appropriées lorsque nous cherchons à rapprocher une taxonomie très peu structurée d'une autre qui l'est davantage. Ces techniques sont néanmoins également utilisables pour des taxonomies très structurées mais, dans ce cas, la richesse structurelle de l'une des deux taxonomies n'est pas exploitée.

Liens de sous-classes uniquement : Les techniques proposées interprètent les liens entre concepts dans les taxonomies traitées exclusivement comme des liens de sous-classes. L'approche est donc inadaptée pour aligner des hiérarchies de classification au sein desquelles les relations entre concepts sont diversifiées (relations d'instanciation, relation partie-de, etc.).

Taxonomies très spécifiques : Les techniques que nous proposons d'exécuter en priorité sont des techniques terminologiques exploitant des mesures de similarité basées sur des

comparaisons de chaînes de caractères. Ces techniques sont bien adaptées lorsque les taxonomies sont des descriptions très fines de domaines d'application car, dans ce cas, des concepts très spécialisés dont le label traduit cette spécialisation sont représentés.

Labels composés de plusieurs mots : L'approche exploite la richesse des labels des concepts. Elle suppose que des concepts proches partagent une partie de leur label. Elle est ainsi plus appropriée pour aligner des taxonomies dont les noms de concepts sont des expressions composées de plusieurs mots car, dans ce cas, ces noms peuvent partager des mots, ce qui peut être révélateur de points communs entre les concepts concernés.

Labels de concepts généraux inclus dans les labels de concepts plus spécifiques : L'approche suppose que très souvent le label d'un concept est construit en reprenant le label du concept qu'il spécialise et en lui ajoutant des qualificatifs permettant de décrire ses spécificités.

3.2 Deux types de relations

Le processus d'alignement génère des mappings 1-1 qui sont des relations de deux types : des relations d'équivalence et des relations de sous-classes.

Relations d'équivalence : Une relation d'équivalence *isEq* est un lien entre un concept dans T_{Source} et un concept dans T_{Cible} dont les noms sont considérés comme étant similaires. Cette similarité recouvre des réalités variées. Il s'agit tout d'abord de relier des termes dont les noms sont rigoureusement identiques syntaxiquement. En effet, les taxonomies auxquelles nous nous intéressons sont très spécialisées et comportent très peu d'homonymes. Il s'agit par ailleurs de relier des termes dont les noms sont des expressions composées de mots qui bien que n'étant pas toujours ordonnées à l'identique, ont la même signification. Il en est ainsi de Pork sausage (liver) et Pork liver sausage. Liver est ici un qualificatif qui peut être soit placé devant le nom qu'il caractérise ou après, en apparaissant entre parenthèses.

Relations de spécialisation : Les relations de spécialisation *isA* sont des liens usuels de sous-classe/super-classe. Quand ce type de lien relie un concept de T_{Source} à un concept de T_{Cible}, son degré de généralité est le même que celui reliant ce super-élément à d'autres sous-éléments de T_{Cible}.

3.3 Une approche basée sur la mesure de similarité $Sim_{LIN\text{-}Like}$

L'ensemble du processus d'alignement est basé sur la mesure de similarité de Lin (Lin, 1998) calculée entre chaque concept de T_{Source} et tous les concepts de T_{Cible}.

$$Sim_{Lin}(x,y) = 2 * \frac{\sum_{t \in tri(x) \cap tri(y)} \log P(t)}{\sum_{t \in tri(y)} \log P(t) + \sum_{t \in tri(y)} \log P(t)}$$

Lin mesure la similarité entre deux éléments, *x* et *y*, en se basant sur le nombre de tri-grammes partagés par les noms de ces deux éléments. Dans la formule, *tri(x)* représente tri-gramme(x), i.e. l'ensemble des tri-grammes de la chaîne de caractères *x* et *P(t)* représente la probabilité d'apparition du tri-gramme *t* dans les termes du corpus. Cette probabilité est

supposée indépendante des autres tri-grammes de la chaîne *x* et est estimée par la fréquence du tri-gramme dans le corpus.

Nous avons adapté cette mesure pour prendre en compte l'importance des mots dans les expressions. Pour traiter ce problème, nous considérons que l'ensemble des tri-grammes communs aux concepts *x* et *y* se partitionnent en deux sous-ensembles *I'* et *INTER*. *I'* contient les tri-grammes communs extraits à partir des mots considérés comme moins importants alors que *INTER* comprend les autres. Un cœfficient dont il est possible de faire varier la valeur permet d'affecter à *I'* un poids dans la formule qui est différent de celui affecté à *INTER*. Si le poids de *INTER* est 2 (cœfficient existant dans la formule de *Lin* s'appliquant à l'ensemble des tri-grammes), celui de *I'* pourra être inférieur à 2. C'est le concepteur du système qui décide de la liste des mots qui, pour le domaine étudié, doivent être considérés comme étant moins importants. Leur définition se fait de manière déclarative.

$$Sim_{Lin\_Like}(x,y) = \frac{2 * \sum_{t \in Inter} \log P(t) + 0,5 * \sum_{t \in I'} \log P(t)}{\sum_{t \in tri(y)} \log P(t) + \sum_{t \in tri(y)} \log P(t)}$$

Etant donné un concept c_S de T_{Source}, cette mesure permet de calculer *MC*, l'ensemble des concepts de T_{Cible} candidats au mapping. *MC* comprend les concepts de la taxonomie cible ayant une forte valeur de similarité avec c_S (seuls les 3 concepts les plus similaires b_1, b_2 et b_3 sont retenus) et les éléments de *Inc*, l'ensemble des concepts de T_{Cible} dont le label est inclus dans le label de c_S. Les techniques successivement appliquées ont alors pour but de sélectionner le concept le plus pertinent de *MC* pour la mise en correspondance, ce concept ne correspondant pas nécessairement à celui qui a la plus forte valeur de similarité, comme nous le montrerons dans ce qui suit.

3.4 Un enchaînement de techniques

Plusieurs techniques sont utilisées : des techniques terminologiques puis structurelles (cf. FIG. 1) après un traitement préalable de normalisation des concepts (remplacement des signes de ponctuation et des symboles spéciaux par des espaces, lemmatisation). Elles sont appliquées séquentiellement de façon à ce que le processus d'alignement soit le plus efficace possible, étant donné la nature des taxonomies sur lesquelles nous travaillons (Kefi et al. 2006). Pour chaque technique, l'objectif est de sélectionner le meilleur concept de T_{Cible} parmi l'ensemble des concepts candidats (dont la mesure de similarité n'est pas nulle).

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 ← MappingCandidates(c_S)
4. **If** TerminologicalMappings(c_S,MC) **then** stop
5. **Else** StructuralMappings(c_S,MC)

FIG.1- *Algorithme d'alignement.*

L'approche combine génération automatique et aide à la découverte de mises en correspondance. Le processus de génération automatique, dit de découverte de mappings probables, permet de découvrir les éléments dont les mappings ont de fortes chances d'être pertinents (Kéfi et al. 2006). Dans un deuxième temps et lorsqu'il est impossible de trouver des mises en correspondance pertinentes de manière totalement automatique, des suggestions de mappings potentiels sont faites à l'utilisateur. Ces mappings correspondent à des concepts entre lesquels l'existence d'une mise en correspondance est suggérée sans que le type exact de relation qui les relie soit systématiquement précisé. Ces deux étapes sont suivies d'une phase de validation, à la charge de l'expert, qui, sur la base des informations qui lui sont communiquées, doit valider et compléter les suggestions soumises par le système.

4 L'exploitation de caractéristiques structurelles

4.1 L'exploitation de la structure de la taxonomie cible : STR_T

Cette technique exploite les concepts candidats à un mapping, (*MC*), avec un concept c_S de T_{Source}. Lorsqu'il n'a pas été possible de générer un mapping probable avec l'un de ces candidats, l'idée consiste à exploiter leur position dans $T_{Cible.}$ Leur proximité dans le graphe est assimilée à une proximité sémantique.

Pour cela, nous identifions le sous-graphe dont la racine est associée à un concept qui n'est pas trop général et tel que ce sous-graphe regroupe un maximum de nœuds de *MC*. Il représente le contexte partagé par un grand nombre de candidats au mapping. Nous faisons l'hypothèse que le concept c_S de T_{Source} peut être mis en correspondance avec un nœud de ce sous-graphe.

La technique STR_T est basée sur le calcul du plus petit ancêtre commun, *LCA*, des concepts de *MC*. A titre d'illustration, FIG. 2 représente le sous-graphe de T_{Cible} comprenant les concepts de $MC = \{b_1 = \text{pork meat tissue}, b_2 = \text{beef connective tissue}, b_3 = \text{beef fat}\} \cup Inc = \{\text{beef}\}$ pour c_s = beef adipose tissue. Le nœud Fresh meat est le *LCA* de tous les concepts de *MC*.

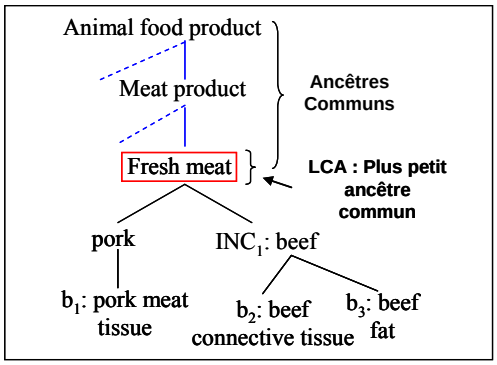

FIG.2 – *Sous-graphe représentant les éléments de MC dans* T_{Cible}

Dans le meilleur des cas, si tous les éléments de *MC* ont le même père dans T_{Cible}, le *LCA* est ce père et le concept c_S est probablement aussi un fils. Dans le cas contraire, si aucun des

éléments de *MC* n'a le même père, le *LCA* est un nœud à un plus haut niveau dans la taxonomie. Trop général, ce *LCA* n'est pas pertinent.

Le *LCA* des éléments de *MC* est d'autant plus haut dans la taxonomie que ses éléments sont éloignés les uns des autres. Nous proposons alors une mesure, la densité relative *DR(Anc*), pour évaluer les sous-graphes regroupant des nœuds d'un sous-ensemble de *MC*. Pour chaque sous-graphe dont la racine est *Anc, le LCA,* et contenant les éléments MC_{Anc} nous calculons sa densité relative *DR(Anc)*. Le sous-graphe dont la densité relative est la plus grande est le plus pertinent. Cette densité prend en compte trois critères : (1) le nombre d'éléments de MC_{Anc}, (2) Sim_{Lin_Like}, la similarité entre les éléments de MC_{Anc} et c_S (3) la distance des éléments de MC_{Anc} à *Anc* en nombre d'arcs.

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

Sur FIG.2, Fresh meat est le *LCA* de quatre candidats au mapping avec une somme des distances de 7 (dist(b_1, Fresh meat) + dist(Inc_1, Fresh meat) + dist(b_2, Fresh meat) + dist(b_3, Fresh meat) = 2+1+2+2). Cependant, beef est l'ancêtre partiel de 3 candidats au mapping {beef, beef connective tissue, beef fat} avec une somme des distances qui est seulement de 2 (dist(Inc_1, beef) = 0). Les résultats donnés par la formule de calcul de *DR* sont présentés FIG.3. La densité relative de beef, *DR*(beef), est la plus élevée. Ainsi, l'ancêtre le plus pertinent est beef.

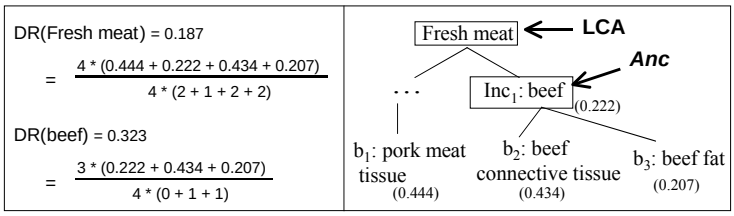

FIG.3 - *Résultats de la densité relative pour les éléments Fresh meat et beef*

L'ancêtre le plus pertinent *Anc* ayant été identifié, nous notons C_{MaxAnc} le nœud ayant la plus forte valeur de similarité dans le sous-graphe dont la racine est *Anc*. Il est considéré comme le candidat au mapping le plus similaire à c_S. L'ancêtre le plus pertinent peut être vu comme un concept définissant le contexte commun à tous ses fils. C'est le contexte dans lequel un grand nombre de candidats au mapping prennent leur sens. Nous faisons l'hypothèse que c_S doit être interprété par rapport à ce même contexte, évitant les mappings avec les concepts dont la valeur de similarité est un peu plus élevée mais qui prennent leur sens dans un autre contexte (exemple : pork meat tissue sur FIG.3). Si C_{MaxAnc} appartient à *Inc*, l'ensemble des concepts dont le label est inclus dans le label de c_S, il est considéré comme un père possible de c_S. Dans le cas contraire, C_{MaxAnc} est considéré comme un frère possible et son père (qui n'est pas nécessairement *Anc*) sera considéré comme un père possible de c_S. Sur la FIG.3, beef est supposé être l'ancêtre le plus pertinent. beef connective tissue est le nœud de MC_{Anc} avec la plus forte valeur de similarité avec c_S (c_S = beef adipose tissue*)*. Ainsi, beef adipose tissue sera considéré comme un frère de beef connective tissue, lié à beef par un lien de sous-classe.

4.2 L'exploitation de la structure de WordNet : STR_W

Les techniques décrites précédemment ne sont pas suffisantes quand les concepts sont sémantiquement proches mais que leur nom est différent. Ainsi, aucune de ces techniques ne permet de rapprocher cantaloupe et watermelon alors que l'interrogation d'une source linguistique, telle que WordNet, peut indiquer que ces concepts sont des sortes de melons et donc sémantiquement très proches. Dans notre approche, l'utilisation des synonymes de WordNet n'est pas suffisante. Nous proposons d'exploiter la structure de WordNet basée sur les liens d'hyperonymie/hyponymie pour trouver, pour chaque concept de T_{Source} non encore mis en correspondance, le concept de T_{Cible} sémantiquement similaire (celui avec lequel il partage des généralisants dans WordNet). Cette approche permet de mettre en correspondance cantaloupe et watermelon qui ne sont pas synonymes mais qui correspondent à deux spécialisations de melon.

L'utilisation de WordNet se fait de la façon suivante. Un expert identifie le nœud, noté $root_A$, qui est le nœud le plus spécialisé de WordNet généralisant tous les concepts du domaine des taxonomies alignées (food dans l'exemple). STR_W recherche ensuite les hypernymes de WordNet de chaque concept de T_{Source} non encore mis en correspondance ainsi que de chaque concept de T_{Cible} (par rapport à tous leurs sens) jusqu'à atteindre $root_A$ ou le terme racine de la hiérarchie WordNet. Par exemple, le résultat de la recherche sur le concept cantaloupe donne les deux ensembles de généralisants suivants qui correspondent à deux sens différents du terme.

Sens 1: cantaloupe → sweet melon → melon → gourd → plant → organism → Living

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

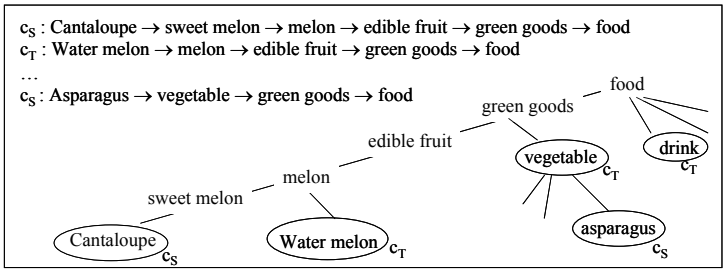

FIG.4 – *Un sous-graphe de T_{WN} liant cantaloupe à watermelon*

Seuls les chemins contenant $root_A$ sont retenus car ils correspondent au seul sens pertinent pour l'application. Un sous-graphe, T_{WN}, composé de l'union des termes et des relations des chemins sélectionnés (cf. FIG. 4) est alors obtenu. Il se compose du nœud racine le plus général de l'application, $root_A$, des feuilles issues des deux taxonomies initiales (cercles sur FIG.4) et des généralisants intermédiaires extraits de WordNet qui peuvent, ou non, appartenir à l'une des deux taxonomies.

Pour chaque concept c_S de T_{Source} pour lesquels un mapping n'a pas encore été proposé, l'objectif est de sélectionner dans T_{WN} le concept de T_{Cible} le plus similaire. Nous utilisons pour cela la mesure de similarité de Wu et Palmer (Wu et Palmer, 1994) selon laquelle la similarité entre deux nœuds c_1 et c_2 est fonction de leur profondeur, depth(c_i), $i \in [1,2]$, i.e.

leur distance à la racine en nombre d'arcs, et de celle de leur plus petit ancêtre commun (*LCA*). Elle est plus précise qu'une mesure basée sur une simple distance des nœuds. En effet, plus la profondeur du *LCA* de deux concepts est importante, plus les deux concepts partagent de caractéristiques communes et plus ils sont proches.

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

Sur FIG.5, si nous recherchons le concept le plus similaire à c_S parmi tous les nœuds de T_{WN} appartenant à T_{Cible} : X_1, X_2, Y et Z, les similarités calculées par la mesure sont, par ordre décroissant : 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).

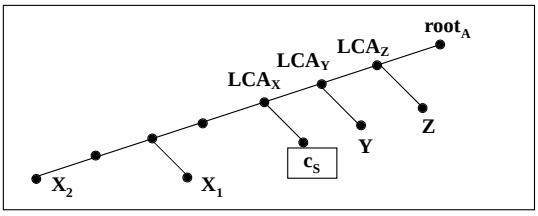

FIG.5 - *Exemple d'arbre T_{WN}*

Notre stratégie d'exploration de l'arbre permet de minimiser le nombre de calculs de valeurs de similarité à effectuer. Cette stratégie a été élaborée à partir d'une analyse des mesures de Wu et Palmer. Du fait du mode de calcul de la mesure, nous savons que le concept le plus similaire d'un nœud c_S est son père, $father(c_S)$. Nous montrons ensuite que la similarité est plus grande entre c_S et un de ses nœuds frères ou un des descendants proches de ce frère, qu'entre c_S et son grand-père, $GF(c_S)$, et ce, jusqu'à une certaine profondeur p du descendant, qu'il est possible de déterminer a priori, pour un élément c_S étant donnée sa profondeur dans l'arbre. En effet, dans tout arbre, le *LCA* d'un nœud et d'un de ses ancêtres est cet ancêtre, i.e., $LCA(c_S, father(c_S)) = father(c_S)$, $LCA(c_S, GF(c_S)) = GF(c_S)$ et le *LCA* d'un nœud et d'un de ses frères ou d'un descendant de son père est ce père, $LCA(c_S, desc(father(c_S))) = father(c_S)$. Si n est la profondeur de c_S dans T_{WN}, ($depth(c_S) = n$), la profondeur du *LCA* de ce nœud et d'un de ses ancêtres sera la profondeur de l'ancêtre considéré, c'est-à-dire la profondeur du nœud c_S moins sa distance en arcs à son ancêtre : $depth(LCA(c_S, father(c_S)))= n-1$ et $depth(LCA(c_S, GF(c_S))) = n-2$.

On peut ainsi calculer la similarité de c_S avec son grand-père, $GF(c_S)$.

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

On peut ensuite calculer la similarité de c_S avec un de ses frères ou un descendant quelconque de son père à une profondeur p.

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

On cherche ensuite à partir de quelle valeur de p, la similarité du grand-père est supérieure à celle d'un descendant du père.

Si $n > 2$,

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

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

Le même calcul peut être effectué pour évaluer à partir de quelle profondeur *p'*, la similarité de l'arrière grand-père doit être prise en compte, et ainsi de suite.

Une fois ces bornes calculées, la stratégie de recherche du terme de T_{Cible} le plus proche d'un élément c_S donné est la suivante : nous testons d'abord si le père de c_S dans T_{WN} est un élément de T_{Cible}. Si c'est le cas, ce père de c_S est le concept qui lui est le plus similaire suivant la mesure de Wu & Palmer. Dans le cas contraire, nous recherchons un nœud de T_{Cible} parmi les descendants du père de c_S, jusqu'à la profondeur *p* (ou égale à *p*). Cette profondeur atteinte, si aucun élément de T_{Cible} n'a été trouvé, nous testons alternativement le grand-père, puis les descendants du père de profondeur (*p+1*), puis les descendants directs du grand-père, et ce jusqu'à atteindre les descendants du père de profondeur p'. Nous testons alors l'arrière grand-père, puis reprenons les tests en alternant chacune des directions, et ainsi de suite. Dès qu'on rencontre un concept de T_{Cible} dans une des directions, on vérifie s'il n'existe pas d'autres éléments de T_{Cible} candidats parmi les descendants directs des derniers nœuds explorés dans les autres directions. Si d'autres candidats existent, on calcule la similarité de ces différents candidats avec c_S et on retient le meilleur, sinon on retient le seul terme de T_{Cible} rencontré, sans qu'aucun calcul de similarité ne soit effectué.

Sur FIG.5, le père de c_S n'appartient pas à T_{Cible}. L'élément c_S est à la profondeur 4, son grand-père à la profondeur 2, la profondeur limite *p* calculée pour c_S est de 5. Aucun descendant du père de c_S à une profondeur égale ou inférieure à 5 n'appartient à T_{Cible}. Après avoir vérifié que le grand-père de c_S n'appartient pas à T_{Cible}, on calcule la profondeur *p'*, égale à 11, associée à l'arrière grand-père de profondeur 1. Les descendants du père de c_S avec une profondeur de 6 sont alors testés. X_1 appartient à T_{Cible}. On cherche alors s'il existe d'autres candidats, fils directs des nœuds déjà explorés et à une profondeur inférieure ou égale à *p'*. Y appartenant aussi à T_{Cible}, on calcule $sim_{W\&P}(c_S,X_1) = 0,6$ et $sim_{W\&P}(c_S,Y) = 0,57$. X_1 est retenu comme l'élément le plus similaire à c_S.

Cette technique permet d'établir des mises en correspondance entre des éléments connus de WordNet, i.e. dont les labels sont des expressions composées généralement de peu de mots. Le type de relation exacte reliant les concepts n'est pas précisé.

4.3 L'exploitation conjointe de la structure des 2 taxonomies : STR_S

A cette étape du processus de découverte de mappings, nous proposons d'appliquer des heuristiques inspirées de celles proposées dans (Melnik et al., 2002, Madhavan et al. 2001). L'idée de base est de faire une proposition de mise en correspondance à partir de l'étude des mappings des nœuds voisins déjà établis. Ainsi, dans l'exemple représenté FIG. 6, le problème est de trouver un mapping pour le terme Apple Cider with 12-14 Brix, fils de Fruit and fruit products dans T_{Source}. Sachant que la majorité des descendants du concept Fruit and fruit

products dans T_{Source} ont été reliés au concept Drink ou à l'une de ses spécialisations dans T_{Cible}, il est vraisemblable que le terme Apple Cider with 12-14 Brix puisse également être rattaché à un élément du sous-arbre de racine Drink. Le problème est donc de déterminer un nœud général dans T_{Cible} similaire à c_S, puis si c_S doit être rattaché à ce nœud général (Drink sur FIG. 6) ou à un nœud plus spécialisé (par exemple Apple juice sur FIG. 6).

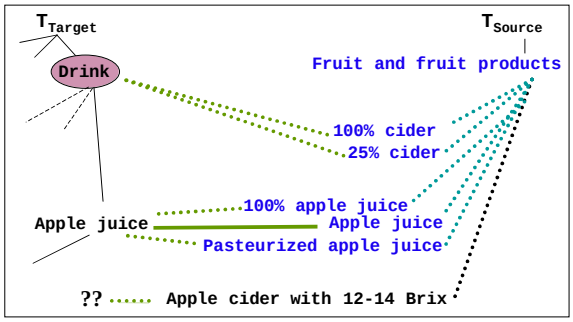

FIG. 6 - *Mappings des frères de Apple cider with 12-24 Brix*

Etant donné un concept c_S feuille de T_{Source}, nous définissons l'ensemble MappingsOfNeighbours (MoN) comme l'ensemble composé des termes de la taxonomie cible auxquels les concepts frères de c_S ont été rattachés par un mapping. Pour chaque élément de MoN, nous mémorisons le nombre de mappings établis avec un frère de c_S. Seuls les éléments de MoN intervenant dans au moins deux mappings sont retenus. Soit CMoN cet ensemble. Les nœuds généraux pertinents retenus sont les nœuds dont les enfants représentent au moins un tiers des éléments de CMoN. Les éléments de CMoN sont présentés à l'expert regroupés par nœuds généraux pertinents et ordonnés par nombre décroissant de mappings établis.

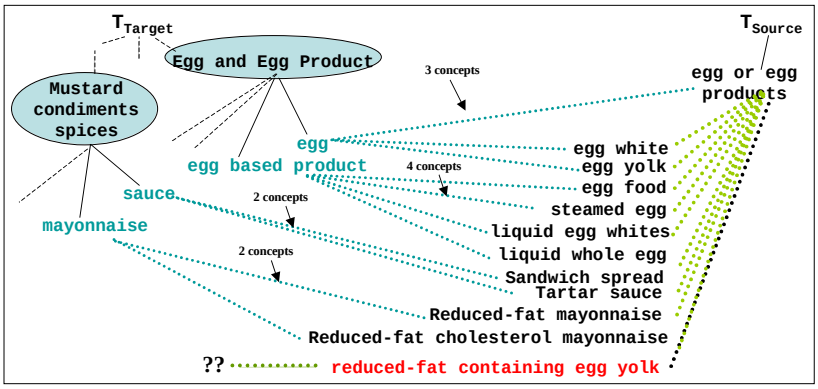

FIG. 7. *Mappings des nœuds frères de reduced-fat containing egg yolk.*

Sur FIG.7, 3 parmi les 11 frères de reduced-fat containing egg yolk ont été mis en correspondence avec egg, 4 avec egg based product, 2 avec sauce, 2 avec mayonnaise. Dans T_{Source}, egg and egg based product ont un père commun Egg and Egg products. Le père de sauce and mayonnaise est Mustard, condiments, spices. Le système effectue deux propositions.

Chacune d'elle est une suggestion d'un mapping soit avec le nœud général (i.e. Egg and Egg product or Mustard, condiment, spices*)* soit avec l'une de ses spécialisations, mais sans toutefois être en mesure de choisir ou de préciser le type de relation exacte reliant le concept.

Si c_S n'est pas une feuille dans T_{Source}, la recherche des éléments de MoN est faite sur les fils de c_S, ses frères et leurs descendants. Un mapping est proposé avec l'élément de CMoN qui est le plus petit ancêtre commun de CMoN.

Cette technique donne des propositions pertinentes lorsque des mappings ont déjà été trouvés pour de nombreux concepts frères ou fils, même si les schémas à apparier sont assez éloignés structurellement.

5 Expérimentations

Les techniques de mapping présentées précédemment sont mises en œuvre dans des modules logiciels indépendants regroupés dans TaxoMap, un prototype développé en Java. Deux types d'expérimentations ont été réalisés. Une première expérimentation a été faite dans le cadre du projet e.dot[1], sur deux taxonomies, Sym'Previus et Com'Base, d'un domaine d'application réel en micro-biologie, le risque bactériologique de contamination d'aliments. Dans un deuxième temps, nous avons appliqué les techniques sur des taxonomies test mises à la disposition de la communauté « Alignement d'ontologies » [2]. Ces taxonomies ne présentent pas la même dissymétrie structurelle et couvrent un domaine beaucoup plus large que celui de notre première expérimentation. Les conditions d'application des techniques ne sont donc pas satisfaites mais notre objectif est de tirer parti des résultats obtenus pour réaliser des améliorations de l'outil et élargir l'approche.

5.1 Résultats dans le domaine de la micro-biologie

Une des tâches réalisées dans le cadre du projet e.dot a été de permettre l'interrogation unifiée des sources d'information Sym'Previus et Com'Base, contenant toutes deux des documents utiles aux experts en microbiologie. L'interrogation devait se faire à partir du moteur d'interrogation MIEL basé uniquement sur l'ontologie Sym'Previus, qui correspond en fait à une hiérarchie de concepts. Une requête formulée uniquement à l'aide du vocabulaire de Sym'Previus doit retourner des documents de Sym'Previus annotés avec les concepts de la requête ou des spécialisations de ceux-ci, et également des documents de Com'Base annotés avec des concepts liés par des mappings à ceux de la requête ou à leurs spécialisations. Ceci n'est possible que si des mappings sont définis entre Sym'Previus et Com'Base. Dans ce processus d'alignement, Sym'Previus est la taxonomie cible et Com'Base, la taxonomie source.

---

[1] E.dot (Entrepôt de Données Ouvert sur la Toile) est un projet de recherche RNTL (2003-2005).

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

L'ontologie de Sym'Previus est une taxonomie composée de 460 concepts reliés par des liens de subsomption. Les concepts sont organisés selon 7 niveaux. Chaque concept de Sym'Previus dispose d'une liste de synonymes et d'une traduction anglaise. Com'Base a été construite par une équipe britannique. Le schéma associé comprend 172 concepts liés par des liens de subsomption. Les concepts sont organisés selon une taxonomie qui n'est pas très structurée. Seuls deux niveaux existent, le premier niveau comprenant uniquement 12 concepts. Les deux taxonomies ne sont pas uniformément développées. Elles ont des parties qui se recouvrent mais ne sont pas structurées de la même façon : une branche peut être très détaillée dans l'une mais pas dans l'autre et inversement. La différence de structure s'explique par le fait qu'elles traduisent des points de vue distincts. Enfin, le niveau de granularité choisi n'est pas le même et le nombre de concepts représentés est différent.

Le domaine d'application couvert par Sym'Previus et de Com'Base est très spécifique. Il est décrit assez finement. Les labels des concepts sont assez longs et comportent généralement plusieurs mots. Souvent, des labels en englobent d'autres, l'ajout de mots à un label d'un concept permettant d'obtenir le label d'un concept plus spécialisé. Les mêmes mots peuvent se retrouver dans plusieurs labels. Ainsi, beaucoup de valeurs de similarité sont non nulles. Le problème qui se pose consiste alors à sélectionner, pour un concept c_S de $T_{Source,}$ le concept le plus pertinent parmi l'ensemble des candidats au mapping (concepts de T_{Cible} avec une mesure de similarité non nulle). Afin d'évaluer notre approche, nous avons demandé à l'expert de fournir les mappings attendus entre ces deux taxonomies, i.e. pour 172 concepts (le nombre de concepts de Com'Base). 44 mappings d'équivalence et 121 mappings de sous-classes ont été proposés par l'expert. 7 concepts n'ont fait l'objet d'aucune proposition (nœuds intermédiaires dans T_{Source}).

Les techniques terminologiques ont généré 101 mappings dont 96 pertinents. Ces techniques ont une précision très importante, supérieure à 90 %, au détriment du rappel (58 %). Les techniques structurelles sont donc très utiles pour compléter les résultats même si les mappings générés sont moins sûrs. En effet, l'enchaînement des trois techniques structurelles proposées permet de trouver 40 mappings sur les 64 restant à identifier (rappel = 40/64 soit 62 %), ce qui n'est pas négligeable. Les résultats obtenus ainsi que la précision de chacune des techniques sont synthétisés dans TAB. 1.

Les techniques structurelles ont été appliquées aux 64 concepts non encore mis en correspondance. Les mappings non trouvés par STR_T correspondent tous à des mappings générés mais non pertinents. Ces mappings non pertinents concernent surtout les concepts dont les labels sont des expressions composées de beaucoup de mots. En effet, cette technique exploite la structure de T_{Cible} mais elle repose aussi beaucoup sur les calculs de similarité entre éléments et ces calculs sont principalement basés sur des comparaisons de chaînes de caractères. Quand les labels sont de longues chaînes de caractères ne comportant qu'un mot référant au concept sous-jacent, parfois très court (exemple : egg ou rice), et que le reste du label ne fait que préciser le concept en employant des mots de T_{Cible} (exemple : rice with chicken protein), la mesure de similarité donne de mauvais résultats. Des erreurs proviennent aussi du fait que seuls quelques mots d'un label sont des mots de T_{Cible} (exemple : home-style salad (reduced calorie mayonnaise with chicken)). Ces problèmes mis à part, la technique exploitant la structure de T_{Cible} a été très utile dans 28 cas sur 42, i.e. dans 66 % des cas.

La technique basée sur WordNet, STR_W, s'est révélée être tout à fait complémentaire des techniques précédemment appliquées. 15 mappings ont été générés. Par exemple, 100% cider a été mis en correspondance avec Drink et Frankfurter avec Sausage. Les 7 concepts pour

lesquels nous n'avons aucun mapping correspondent soit à des acronymes (exemple : TSB), soit à des mots techniques (exemple : Egyptian Kofta) ou des mots trop longs non reconnus de WordNet. Parmi les suggestions erronées, on peut citer, par exemple, Lamb qui a été mis en correspondence avec Meat alors que l'expert proposait de le relier à Sheep, un concept plus spécifique.

Techniques structurelles	# concepts étudiés	# mappings proposés	# mappings confirmés	# mappings non trouvés	Précision
Exploitation de la structure de T_{Cible}	64	42	28	22	66 %
Exploitation de la structure de WordNet	22	15	9	7	60 %
Exploitation de la structure des 2 taxonomies	7	5	3	2	60 %

TAB. 1- *Nombre de mappings obtenus par technique structurelle*

Nous avons appliqué la dernière technique sur les 7 concepts restant à apparier. Deux concepts, TSB et Phosphate buffer, n'ont pas été mis en correspondance car ils n'avaient pas assez de frères pour que la technique soit applicable. Pour les 5 autres concepts, des propositions pertinentes ont été faites, comme lier Tampeh, Brocoli, Pecan and Pecan nuts avec Fresh Fruits and Vegetables, (pour Pecan and Pecan nuts l'expert avait proposé un mapping avec un autre fils de Vegetable, Dried Fruits and Vegetables). Pour le dernier concept, Egyptian Kofta, le système propose 3 directions d'appariement, la première avec Fresh meat, la deuxième avec Meat-based product (ce qui correspond à la proposition de l'expert) et la troisième avec Poultry.

5.2 Application à d'autres taxonomies

Des taxonomies mises à disposition de la communauté « Ontology matching » ont été testées. Les résultats ont montré l'efficacité de notre approche et nous ont donné des idées d'améliorations pour élargir le champ des taxonomies alignées. En effet, les caractéristiques des taxonomies tests sont différentes des caractéristiques de celles qui ont motivé notre approche. Les résultats de ces tests sont synthétisés dans TAB. 2.

5.2.1 Alignement de taxonomies portant sur la russie

Les deux taxonomies alignées, Russia-A et Russia-B, décrivent la Russie du point de vue de sa géographie et de ses monuments[3]. Elles ont approximativement le même nombre de concepts (300) et la même profondeur (7 niveaux). Les labels des concepts ne comportent bien souvent qu'un seul mot. Ces caractéristiques les rendent très différentes des taxonomies des premières expérimentations et vont peser différemment sur le processus d'alignement.

[3] http://www.atl.external.lmco.com/projects/ontology/i3con.html

Taxomap a généré 96 mappings d'équivalence sur les 103 attendus. Les 7 concepts sans mapping auraient dû être mis en correspondance avec des concepts sémantiquement équivalents mais de label différent. Par ailleurs, TaxoMap propose 29 mappings de sous-classes supplémentaires. Ces résultats sont satisfaisants car beaucoup de mappings attendus sont retrouvés. Cependant, les caractéristiques des taxonomies de permettent pas d'exploiter les forces de l'approche. Lorsque les labels n'ont qu'un mot, ce mot est différent d'un concept à un autre. Ainsi, les labels de concepts liés par un lien de sous-classe partagent rarement des mots. Durant le processus d'alignement, un concept c_S de T_{Source} a souvent un nombre très limité de candidats au mapping dans T_{Cible}. Notre approche conçue pour sélectionner le concept le plus pertinent parmi un ensemble de candidats, avec l'hypothèse que le nombre de candidats est le plus souvent supérieur à 3, n'est dans ce cas pas très opérationnelle.

Ceci étant dit, de bons résultats sont obtenus lorsqu'il y a au moins trois candidats au mapping (assez rare). Par exemple, la technique basée sur l'inclusion de labels appliquée lorsque le label inclus correspond à un concept qui a la valeur de similarité la plus grande permet d'identifier une quinzaine de mappings pertinents, tels que Azov_sea ou black_sea is-a sea, capital_city is-a city, cathedral_of_sophia is-a cathedral, ou monetary_unit is-a unit. Ces mappings ne sont pas des mappings attendus fournis avec les taxonomies tests car seuls des mappings d'équivalence ont été donnés, mais ils sont quand même pertinents. D'un autre côté, lorsque le seul candidat au mapping est un concept dont le label est inclus dans le label de c_S, des mappings erronés sont générés, comme North_America *isA* North ou Easter *isA* East.

La technique basée sur WordNet, STR_W, n'est pas adaptée pour aligner ces taxonomies du fait du domaine d'application qui est beaucoup trop large. La technique construit un sous-arbre à partir de tous les nœuds hypernymes de WordNet jusqu'à atteindre le nœud le plus général de l'application. Dans le cas d'un domaine très large, le concept le plus général est un nœud placé très haut dans la hiérarchie WordNet, si ce n'est le nœud racine. T_{WN} est donc très gros. Il mêle des sens de termes différents et conduit à générer des mappings qui ne sont absolument pas pertinents. Des améliorations seraient possibles si plusieurs sous-arbres étaient construits, un par sous-domaine traité dans la taxonomie en supposant que les différents sous-domaines puissent être identifiés.

La dernière technique structurelle, STR_S, exploitant les mappings des nœuds frères du concept étudié dans T_{Source} n'est pas adaptée non plus car les nœuds de T_{Source} ont très peu de nœuds frères.

Cette expérimentation nous a permis de réfléchir à des adaptations possibles de TaxoMap afin d'obtenir des résultats plus pertinents. Actuellement, il n'y a que un ou deux candidats au mapping, le concept ayant la plus grande valeur de similarité est retenu. Ainsi, bus a été considéré comme un frère de foreign_business_person et Chechnya a été considéré comme un frère de Check alors que les deux valeurs de similarités étaient très faibles. Cela montre la nécessité de rejeter des mappings avec une valeur de similarité trop faible (inférieure à un certain seuil) ou d'essayer de les confirmer par application d'une autre technique. L'interrogation de WordNet pourrait, par exemple, permettre d'effectuer une telle confirmation.

5.2.2 Alignement de taxonomies correspondant à des catalogues de cours

Les deux taxonomies alignées couvrent un domaine également important. Elles contiennent des informations portant sur des cours de deux universités, Cornwell et

Washington[4]. Elles ont, là aussi, le même nombre de concepts (150) et la même profondeur (4 niveaux). Contrairement aux taxonomies précédentes, ici les concepts ont des labels composés de plusieurs mots. Beaucoup de mots interviennent dans plusieurs labels. Les candidats à un mapping sont donc plus nombreux que dans le cas précédent. Nos techniques peuvent aider à sélectionner le plus pertinent d'entre eux.

50 mappings d'équivalence étaient proposés avec les taxonomies. TaxoMap en a reconnu 45. Dans 35 cas, le concept cible est considéré comme étant équivalent, dans 10 autres cas, le concept cible est considéré comme le concept frère le plus probable. 77 mappings supplémentaires ont été proposés par notre système. Nous estimons personnellement que 52 d'entre eux sont corrects.

Les techniques générant des mappings probables ont une très bonne précision. L'inclusion entre labels nous permet d'identifier 9 relations de sous-classes et toutes sont pertinentes (par exemple : Applied_Mathematics *isA* Mathematics, French_Linguistics *isA* Linguistics). La technique basée sur la mesure de similarité significativement plus élevée que les autres permet de générer 15 mappings pertinents supplémentaires parmi 16 proposés, par exemple : Political_Science est considéré comme un frère de Political_Theory, International_Studies_Jewish_Studies est considéré comme un frère de Program_of_Jewish Study. La grande précision de cette technique montre que ces mappings sont sûrs.

Dans cette experience, les techniques structurelles ont souvent été utilisées pour découvrir des mappings supplémentaires. La technique exécutée sur $T_{Cible,}$ STR_T, est appliquée 43 fois et fournit 24 mappings pertinents. Par exemple, Ancient_and_Medieval_History est lié à Medieval_Renaissance et à Early_Modern_European_History, deux sous-classes de History, Biology est proche de Plant_Biology et de Microbiology, ayant comme père commun le concept Decision of Biological Science. La précision de la technique n'est toutefois pas aussi bonne ici que lors de la première expérimentation pour plusieurs raisons : le domaine couvert est plus important, les labels sont moins précis, la sémantique des concepts n'est pas seulement donnée par les labels qui ne sont pas toujours très signifiants mais aussi par leur position dans la hiérarchie, les techniques sont appliquées séquentiellement. Du fait que le domaine couvert est plus important, les concepts sont plus généraux et doivent être interprétés dans le contexte de la hiérarchie. Par exemple, Literature et Language_course sont plutôt des concepts généraux mais dans le cadre de la taxonomie Course Catalog, ils doivent être interprétés dans le contexte de Near_Eastern_Studies. Notre approche n'exploite pas simultanément toutes les informations structurelles ce qui explique que des mappings erronés soient générés. Par exemple, nous pourrions souhaiter lier Slavic_languages_and_Literatures à Russian_Language qui appartient à l'ensemble des candidats au mapping. Au contraire, TaxoMap propose de rapprocher ce concept de Literature et de Language_Course en établissant une relation de sous-classe avec leur père commun Near_Eastern_Studies. Ceci n'a aucun sens. Enfin, comme pour le cas précédent, le domaine d'application est trop large et les concepts de T_{Source} ont trop peu de fils pour STR_W et STR_S soient opérationnelles.

Ces expérimentations nous ont montré quelles étaient les forces et les faiblesses de notre approche et comment il était possible de l'améliorer. L'approche convient très bien pour des taxonomies très fines contenant seulement des liens de sous-classes. Elle est moins bien adaptée à l'alignement de taxonomies générales modélisant implicitement d'autres relations

[4] http://anhai.cs.uiuc.edu/archive/domains/course_catalog.html

telles que des relations partie-de, instance-de, etc. Toutefois, quelles que soient les taxonomies alignées, l'approche est capable de retrouver presque tous les mappings d'équivalence attendus. Le point fort de TaxoMap est de proposer en plus d'autres mappings pertinents (+ 29 concernant les taxonomies sur la russie, + 52 concernant les taxonomies sur les cours, en plus de ceux qui étaient attendus). Certains mappings ont une grande précision et sont sûrs (générés par les techniques terminologiques). D'autres (générés par les techniques structurelles) sont moins sûrs (précision faible) et doivent être validés par un expert humain mais, si l'implication d'un expert est possible, l'approche est très intéressante car cela permet d'obtenir beaucoup plus de mappings.

	Course catalog		Russia	
	T_T	T_S	T_T	T_S
	Cornwell	Washington	Russia-A	Russia-B
# Concepts	176	167	372	310
# mappings attendus	50		103	
# mappings pertinents trouvés	45		96	
# mappings supplémentaires pertinents	52		29	

TAB. 2. - *Nombre de mappings générés à partir des taxonomies tests*

6 Conclusion et perspectives

Notre objectif est de construire un système qui génère automatiquement des mappings entre deux taxonomies. Nous avons implémenté un prototype, TaxoMap, qui utilise une approche exploitant la syntaxe des noms des concepts, la structure des taxonomies et un thesaurus, WordNet, pour établir automatiquement des mappings entre des éléments de deux schémas. Les mappings se distinguent selon leur plausibilité. Seuls les mappings probables sont découverts automatiquement. Lorsque le système ne peut pas identifier de mappings probables, il aide l'utilisateur à le faire en lui indiquant un ensemble de mappings possibles.

Le contexte dans lequel nous nous plaçons ne nous permet pas de retrouver des structures identiques dans les deux taxonomies alignées. Nous proposons alors d'autres façons d'exploiter la structure : exploitation de la structure de la taxonomie cible uniquement, exploitation de la structure de la hiérarchie hyperonymie/hyponymie de WordNet, exploitation de la structure des 2 taxonomies combinées avec l'exploitation de mappings préalablement identifiés. Ces techniques sont originales car elles se distinguent de la recherche de similarités structurelles. Elles permettent de faire des suggestions de mappings moins sûrs que ceux générés par les techniques terminologiques, ce qui explique que les techniques terminologiques sont appliquées en premier. Néanmoins, il s'agit d'un bon complément comme le montrent les expérimentations.

TaxoMap a été testé dans le cadre du projet e.dot avec deux taxonomies réelles dans le domaine de la microbiologie et également sur des taxonomies servant de tests aux travaux réalisés dans le domaine.

Ce travail est original parce qu'il pose le problème de la génération de mappings lorsque l'on dispose de peu de critères. En effet, les schémas sont des taxonomies au sein desquelles les concepts sont définis surtout par rapport à la terminologie du nom qui leur est associé et dont les niveaux de profondeur peuvent ne pas être très nombreux. Les expériences réalisées montrent qu'une approche enchaînant des techniques particulières adaptées au contexte décrit, terminologiques et structurelles, peut donner de bons résultats.

Ce travail va se poursuivre par la construction d'une boite à outils proposant nos techniques ainsi que d'autres, chacune étant adaptée à des taxonomies ou ontologies particulières. Nos techniques conviennent bien pour aligner des taxonomies structurellement dissymétriques couvrant un domaine relativement petit, décrit finement à l'aide de concepts dont les labels sont des expressions complexes. D'autres techniques seraient utiles pour élargir notre approche. Elles pourraient être adaptées des techniques présentées dans cet article ou être issues de modifications plus importantes afin de prendre en compte, entre autres, la couverture du domaine d'application, la complexité des expressions utilisées pour nommer les concepts, la profondeur des taxonomies, le nombre de termes équivalents, la volonté d'implication de l'expert. Les adaptations nécessaires sont de deux types : adaptations des techniques elles-mêmes ou de la façon de les utiliser (application séquentielle/combinaison).

Remerciements

Nous remercions Hassen Kefi et Ahlem Slimi pour leur contribution à ce travail.

References

Bach, T.-L., Dieng-Kuntz, R., Gandon, F. (2004). *On Ontology Matching Problems for building a Corporate Semantic Web in a Multi-Communities Organization*. In the proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, April 14-17.

Do, H. H., Rahm, E. (2001). *COMA – A system for flexible combination of schema matching approaches*. In the proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001), pp. 610-621.

Doan, A., Madhavan, J., Domingos, P., Halevy, A. (2002). *Learning to map between ontologies on the semantic web*. In the proceedings of the eleventh International WWW Conference. Haway, USA, pp. 662-673.

Giunchiglia, F., Shvaiko, P. (2003). *Semantic Matching*. Knowledge Engineering Review. 18(3):265-280.

Kalfoglou, Y., Hu, B. (2005). *CROSI Mapping System (CMS) Results of the 2005 Ontology Alignment Contest*. Integrating Ontologies Workshop, In the proceedings of the third International Conference on Knowledge Capture (K-Cap'05). Banff, Canada, October 2-5, pp. 77-84.

Kalfoglou, Y., Schorlemmer, M. (2003). *Ontology mapping: the state of the art*. Knowledge Engineering Review, 18(1):1-31.

Kéfi, H. (2006). *Ontologies et aide à l'utilisateur pour l'interrogation de sources multiples et hétérogènes*. Ph.D Thesis. Université Paris Sud, Mars 2006.

Kéfi, H., Safar, B., Reynaud, C. (2006). *Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes*. 15ème Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (RFIA'06), 25-27 Janvier, Tours.

Lin, D. (1998). *An Information-Theoretic Definition of Similarity*. In proceedings of the Fifteenth International Conference on Machine Learning (ICML-98), Madison, Wisconsin USA, July 24-27, pp. 296-304.

Madhavan, J., Bernstein, P. A., Rahm, E. (2001). *Generic matching with Cupid*. International Journal of Very Large Data Bases (VLDB), 10(4): 334-350.

Maedche, A., Staab, S. (2002). *Measuring similarity between Ontologies*. In proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), Madrid, Spain, October 1-4, LNCS/LNAI 2473, Springer, pp. 251-263.

Melnik, S., Garcia-Molina, H., Rahm, E. (2002). *Similarity Flooding: A versatile Graph Matching Algorithm and its application to schema matching*. Proceedings of the 18th International Conference on Data Engineering (ICDE). San Jose CA, USA, pp. 117-128.

Miller, G. A. (1995). *WordNet: A lexical Database for English*. Communications of the ACM, 38(11): 39-45.

Noy, N. F., Musen, M. A. (2001). *Anchor-Prompt: Using non-local context for semantic matching*. Workshop on Ontologies and Information Sharing, International Joint Conferences on Artificial Intelligence (IJCAI-01), Seattle, Washington, USA.

Shvaiko, P., Euzenat, J. (2004). *A survey of Schema-based Matching Approaches*. Technical Report DIT-04-087. Informatica e Telecomunicazioni, University of Trento.

Wu, Z., Palmer, M. (1994). *Verb semantics and lexical selection*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-94). Las Cruces, pp. 133-138.

Summary

The aim of this paper is to allow a uniform access to documents belonging to a same application domain. In particular, it allows to access additional sources from a Web portal without modifying its querying interface. We assume retrieval of documents is supported by taxonomies and focus on alignment of these heterogeneous taxonomies. Original and specific structural alignment techniques suitable with a dissymmetry in the structure of the mapped taxonomies are presented. Experimental results on both real-world and test taxonomies are given and analyzed.