Techniques structurelles pour l'alignement de taxonomies sur le Web

Hassen Kefi , Chantal Reynaud
Brigitte Safar

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

Résumé. Ce papier porte sur la génération de mappings pour l'alignement de taxonomies du Web. L'objectif est de permettre un accès unifié aux documents d'un même domaine d'application. La recherche de documents s'appuie sur des taxonomies. 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. Ce papier présente les techniques structurelles de la méthode d'alignement que nous avons développée, des techniques qui s'appliquent en présence d'une dissymétrie dans la structure des taxonomies comparées. Des expérimentations ont été effectuées avec le prototype implémenté, TaxoMap.

1 Introduction

La recherche de documents pertinents sur le web est une tâche encore souvent laborieuse. Le Web sémantique devrait faciliter cette recherche en réalisant un appariement sémantique entre la requête de l'utilisateur et les documents indexés. Les techniques d'alignement des schémas de méta-données ou des ontologies sont au cœur du processus.

Notre travail porte sur de telles techniques, utilisables dans le contexte du Web. L'objectif est de permettre un accès unifié via le Web aux documents d'un même domaine d'application. La recherche de documents s'appuie sur des taxonomies de termes plus ou moins structurées. 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. Ce papier présente les techniques structurelles de la méthode d'alignement que nous avons développée, des techniques difficiles à appliquer a priori dans le contexte dans lequel on se situe car, si la taxonomie de concepts d'un portail Web est en général bien structurée, celle des autres documents accessibles ne l'est pas toujours. Les techniques que nous proposons s'appliquent donc en présence d'une dissymétrie dans la structure des taxonomies comparées. Ces techniques font partie d'une approche générique d'alignement de taxonomies mise en œuvre au travers d'un processus semi-automatique. Dans une première étape, des mises en correspondance dites probables sont automatiquement découvertes. Dans une seconde étape, des suggestions de mappings sont faites au concepteur. La découverte de mappings peut être vue comme un assemblage de techniques variées, appliquées dans un ordre bien défini : terminologiques, structurelles et sémantiques. Les techniques terminologiques, basées principalement sur des comparaisons de chaînes de caractères, sont appliquées en priorité car elles sont les plus à

même de fournir des mappings probables. Elles exploitent toute la richesse des noms des concepts. Même si elles sont efficaces, les techniques terminologiques ne peuvent cependant pas trouver l'ensemble des rapprochements possibles. Le système fait alors appel à des techniques structurelles et sémantiques[1]. Ce papier porte sur ces techniques.

Les techniques structurelles permettent de générer des mises en correspondance supplémentaires, moins sûres que celles générées par les techniques terminologiques et qui nécessitent d'être validées. 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 l'approche d'alignement au sein de laquelle s'insèrent les techniques présentées. La section 3 présente successivement les trois techniques structurelles mises en œuvre. La section 4 porte sur les expérimentations réalisées. En section 5, des travaux proches sont cités et nous discutons, à la lumière de ces travaux, les caractéristiques des techniques retenues dans notre approche. Enfin, nous concluons.

2 Approche

L'objectif du processus d'alignement est de générer, le plus automatiquement possible, des appariements sur des taxonomies. Les critères utilisables pour déduire une mise en correspondance sont restreints. En effet, dans une taxonomie, un concept est uniquement défini par le label qui lui est associé (expression qui peut être composée de plusieurs mots) et par les relations de subsomption qui le relient à d'autres concepts.

Une taxonomie est un ensemble de concepts reliés par des relations *is-a*, représentée par des graphes acycliques. Les concepts sont représentés par des nœuds du graphe connectés par les liens orientés correspondant aux relations *is-a*.

Etant donné deux taxonomies, il s'agit de mettre en correspondance les éléments de l'une, appelée taxonomie source, avec les éléments de l'autre, appelée taxonomie cible. Le processus est orienté d'une taxonomie source vers une taxonomie cible. Les mappings à déterminer sont supposées être des relations de type 1:1. Le processus d'alignement a pour objectif de générer deux types de relations : des relations d'équivalence et des relations de spécialisation.

2.1 Deux types de relations

2.1.1 Relations d'équivalence

Une relation d'équivalence *is-equivalent* est un lien entre un élément d'une taxonomie source, $\mathbf{T_{source}}$, et un élément d'une taxonomie cible, $\mathbf{T_{Cible}}$, dont les noms sont 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 spécifiques à des domaines d'application ; il n'existe

[1] Toutes les techniques présentées exploitent la structure de modèles. Parmi celles-ci, l'une exploite conjointement la structure et les relations sémantiques de WordNet. C'est une technique à la fois structurelle et sémantique.

que 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.

2.1.2 Relations de spécialisation

Les relations de spécialisation sont les liens usuels *is-a* sous-classe/super-classe. Quand ils relient un élément de la taxonomie source à un super-élément de la taxonomie cible, le degré de généralité du lien est supposé être le même que dans le lien *is-a* reliant ce super-élément à d'autres sous-éléments dans la taxonomie cible. Ainsi, *Asparagus* de la taxonomie source pourra être relié à *Fresh fruit and vegetables* de la taxonomie cible tout comme *Carrots*, un autre terme de celle-ci.

2.2 Un assemblage de techniques

La découverte de mappings repose sur des techniques variées : terminologiques, structurelles et sémantiques. Ces différentes techniques sont composées de façon à rendre le processus de génération des mappings le plus efficace possible (Kefi et al. (2006)). Les techniques terminologiques sont appliquées en priorité. Elles permettent de générer les mappings les plus probables en exploitant toute la richesse des labels des concepts. Les techniques structurelles et sémantiques permettent de trouver des mappings supplémentaires, potentiellement vrais, lorsque l'exploitation des chaînes de caractères ne suffit pas (cf. FIG. 1). L'avantage d'une telle approche est de fournir une catégorisation des mappings suivant la façon dont ils ont été obtenus. Ceci est important aux yeux de l'expert puisque chaque ensemble de mappings n'a pas la même vraisemblance. Une telle catégorisation peut accélérer le processus de validation.

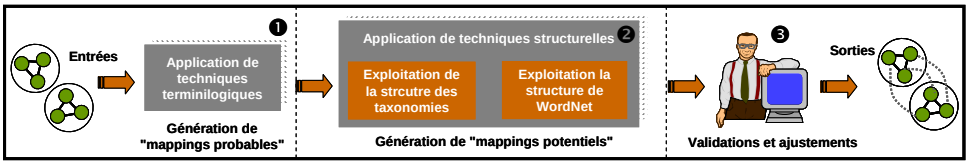

FIG. 1 - *Processus général d'alignement de taxonomies*

3 Exploitation de la structure des représentations

Les trois techniques présentées dans cette section sont utilisées quand des mappings probables n'ont pu être identifiés. Elles exploitent la structure de différentes représentations. La structure de représentation des connaissances la plus riche est supposée être celle de la taxonomie cible et la première technique appliquée s'appuie sur elle. En second, nous proposons d'utiliser une ressource externe, WordNet (Miller (1995)), et d'exploiter sa structure et ses relations sémantiques. 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 taxonomie de la source peut être très peu structurée.

L'objectif est de rattacher un élément e_s de la taxonomie source à un élément e_c de la taxonomie cible. Les mappings générés sont essentiellement des mappings de spécialisation. Ils s'appuient sur le calcul préalable de la similarité de cet élément e_s à tous les termes de la taxonomie cible. La mesure de similarité utilisée est celle de Lin (Lin (1998)) qui compare les chaînes de caractères. Elle a été adaptée pour prendre en compte l'importance des mots dans les expressions.

3.1 L'exploitation de la structure de la taxonomie cible

Dans cette première technique, nous travaillons sur *MC*, l'ensemble des termes candidats à un mapping avec l'élément e_S. Ces termes candidats ont été identifiés à partir du calcul de similarité. Ce sont les termes de $\mathbf{T_{Cible}}$ dont le nom est inclus dans le nom de e_S (INC) ou ceux qui ont une forte similarité avec e_S (seuls les 3 éléments les plus similaires sont retenus). Lorsqu'il n'a pas été possible de déduire un mapping probable avec l'un de ces éléments, l'idée consiste à exploiter leur position dans $\mathbf{T_{Cible}}$. Le sous-graphe représentant les éléments de *MC* au sein de $\mathbf{T_{Cible}}$ est analysé. Dans le meilleur des cas, si tous les éléments de *MC* ont le même père dans $\mathbf{T_{Cible}}$, l'élément e_S considéré a aussi probablement le même père. Dans le cas contraire, si tous les éléments de *MC* n'ont pas le même père, nous cherchons leur plus petit ancêtre commun (Lowest Common Ancestor, LCA). Ainsi, FIG. 2 représente le sous-graphe de $\mathbf{T_{Cible}}$ représentant les éléments de $MC = \{b_1, b_2, b_3\} \cup \{beef\}$ pour e_s = *beef adipose tissue* et INC = {*beef*}. Sur cette figure, l'élément *Fresh meat* représente le plus petit ancêtre commun à tous les éléments de *MC*. Si cet ancêtre commun est un nœud très haut placé dans la taxonomie, il ne sera pas très significatif car trop général.

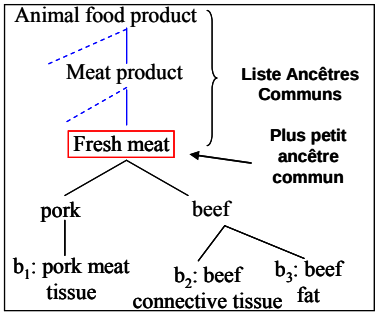

FIG. 2 - *Sous graphe représentant les éléments de MC au sein de* $T_{Cible.}$

Une fois le sous-graphe représentant les éléments de *MC* construit et leur plus petit généralisant au sein de $\mathbf{T_{Cible}}$ identifié, l'idée consiste à rechercher l'élément le plus pertinent qui pourrait être apparié à e_S dans ce sous-graphe. Pour obtenir des suggestions d'appariement les plus pertinentes, nous recherchons des ancêtres partiels, c'est-à-dire des nœuds qui sont les ancêtres d'un sous-ensemble d'éléments de *MC*. Pour chaque sous-ensemble d'éléments de *MC* partageant un ancêtre partiel commun, nous construisons le sous-graphe correspondant qui a pour racine l'ancêtre partiel considéré. Les sous-graphes sont construits comme suit. Nous identifions le plus petit ancêtre commun de chaque paire d'éléments de *MC* et nous étendons l'ensemble considéré élément par élément. Pour un sous-graphe dont la racine est l'ancêtre partiel *Anc*, nous calculons la distance relative *DR(Anc)*

entre les éléments de *MC* correspondant aux nœuds de ce sous-graphe (MC_{Anc}). L'intuition de la formule proposée pour calculer *DR(Anc)* est de tenir compte des trois critères suivants :

- le nombre d'éléments de *MC* dont l'élément *Anc* est l'ancêtre,
- la distance des éléments de *MC* à *Anc* en nombre d'arcs,
- la similarité terminologique des éléments de *MC* à e_S.

$$DR(Anc) = \frac{|MC| * \sum_{ec \in MCAnc} dist(e_c, Anc)}{|MC_{Anc}| * \sum_{ec \in MCAnc} Sim_{LIN\text{-}Like}(e_s, e_c)}$$

Le sous-graphe dont la valeur de *DR(Anc)* est la plus faible est considéré comme le plus pertinent. La distance relative d'un ancêtre partiel *Anc* est d'autant plus faible que :

- *Anc* est l'ancêtre d'un plus grand nombre d'éléments au sein de *MC* ($|MC_{Anc}|$),
- sa distance à ses descendants dans *MC* est faible ($\sum e_c \in MC_{Anc}$ dist(e_c, *Anc*)),
- *Anc* est l'ancêtre d'éléments très similaires de e_S ($(\sum_{ec \in MCAnc} Sim_{LIN\text{-}Like}(e_s, e_c))$.

Dans FIG. 2, l'élément *Fresh meat* est le plus petit ancêtre commun des quatre candidats au mapping, avec une distance de 7 ({(1) + (2)} + {(3) + (4)} + { (3) + (5)} +{(3)}) ce qui fait un résultat de (2+2+2+1). Mais un autre sous-graphe de racine *beef* et regroupant trois des candidats au mapping {*beef, beef connective tissue, beef fat*} peut être construit avec une distance égale à 2 ({(4)} + {(5)}). Ainsi, en appliquant la formule de calcul de distance relative, nous obtenons les résultats illustrés FIG. 3.

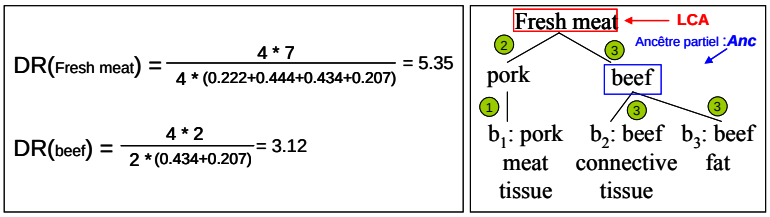

FIG. 3 - *Résultats du calcul de la distance relative pour les éléments Fresh meat et beef.*

Etant donné que la distance relative DR(*beef*) est la plus faible, le sous-graphe le plus pertinent est donc celui dont la racine est *beef*. Dans ce sous-graphe, le nœud qui a la valeur de similarité la plus élevée est noté C_{MaxAnc}. Il est considéré comme le candidat au mapping le plus proche de e_S. Si C_{MaxAnc} appartient à l'ensemble des termes dont le nom est inclus dans celui de e_S, il est considéré comme un père possible de e_S dans $\mathbf{T_{Source}}$. Dans le cas contraire, C_{MaxAnc} est considéré comme un frère possible et son père (qui n'est pas forcément le nœud racine du sous-graphe) est proposé comme un père possible de e_S. Sur l'exemple précédent, le groupement représenté par le sous-graphe dont la racine est *beef* est donc jugé plus pertinent et le nœud de ce sous-graphe qui a la valeur de similarité la plus élevée, C_{MaxAnc} : *beef connective tissue*, est proposé comme un frère de *beef adipose tissue* qui sera relié à l'élément *beef* dans $\mathbf{T_{Cible}}$ par une relation de spécialisation.

3.2 L'exploitation de la structure et des relations sémantiques de WordNet

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 melons. Dans notre approche, l'utilisation des synonymes de WordNet n'est pas suffisante. Nous proposons de combiner l'exploitation des relations sémantiques de WordNet avec la structure de la hiérarchie de WordNet afin de trouver, pour chaque élément de la taxonomie source, de quels éléments de la taxonomie cible il peut être sémantiquement proche (ceux avec lesquels il partage des généralisants dans WordNet). Cela permet, par exemple, de rapprocher *cantaloupe* et *watermelon* qui ne sont pas synonymes mais qui sont deux spécialisations du concept *melon*.

L'utilisation de WordNet s'effectue en deux étapes. La première étape consiste à construire un sous-arbre (appelé S_{WN}) à partir de l'ensemble des généralisants dans WordNet de chacun des éléments des deux taxonomies $\mathbf{T_{Cible}}$ et $\mathbf{T_{source}}$ que l'on cherche à rapprocher. La seconde étape consiste à utiliser une mesure de similarité (Wu et Palmer (1994)) pour identifier, au sein de la taxonomie commune S_{WN}, l'élément de $\mathbf{T_{Cible}}$ le plus proche de l'élément de $\mathbf{T_{source}}$ considéré,. Pour construire S_{WN}, nous interrogeons WordNet pour chaque élément de chacune des deux taxonomies. Pour chaque élément, et pour chacun de ses sens (chaque synset auquel il appartient), nous extrayons l'ensemble de ses généralisants jusqu'à atteindre le concept le plus général pour l'application considérée. Par exemple, le résultat de la recherche sur le terme *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 thing*
- Sens 2 : *Cantaloupe → sweet melon → melon → edible fruit → green goods → food*

Seul est conservé le sens qui contient le concept racine de l'application étudiée (sens 2 dans l'exemple car il contient *food*). Les généralisants de l'élément sont intégrés pour construire le sous-arbre de WordNet pertinent pour l'application, i.e., dont la racine est le terme le plus général de l'application. Les feuilles sont les termes issus des deux taxonomies initiales (dans des ovales sur FIG. 4). Les généralisants intermédiaires sont extraits de WordNet mais ils peuvent également appartenir à l'une des deux taxonomies.

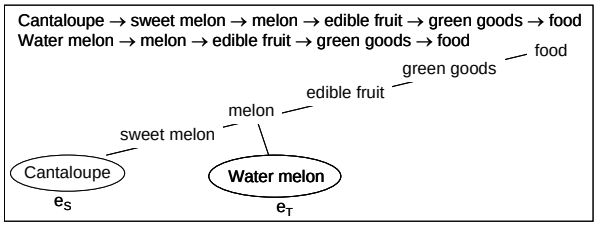

FIG. 4 - *Un sous-graphe de S_{WN} reliant cantaloupe et water melon dont la racine est food.*

Une fois S_{WN} construit, nous évaluons la similarité sémantique entre deux nœuds c_1 et c_2 de l'arbre en utilisant la mesure de similarité définie dans (Wu et Palmer (1994)). Wu et Palmer définissent la similarité sémantique entre deux concepts c_1 et c_2 en fonction de leur profondeur, $prof(c_i)$, $i \in [1,2]$, i.e. leur distance à la racine en nombre d'arcs et en fonction de la profondeur de leur plus petit ancêtre commun (LCA) :

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

Cette mesure est plus précise qu'une mesure basée sur une simple distance. 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. Ainsi, sur l'exemple FIG. 5, si l'on recherche dans S_{WN} le terme le plus proche de l'élément de la taxonomie source e_S parmi les nœuds de T_{cible}, X_1, X_2, Y et Z, les similarités calculées par la mesure sont, par ordre décroissant : $sim_{W\&P}(e_S,X_1) > sim_{W\&P}(e_S,Y) > sim_{W\&P}(e_S,X_2) > sim_{W\&P}(e_S,Z)$.

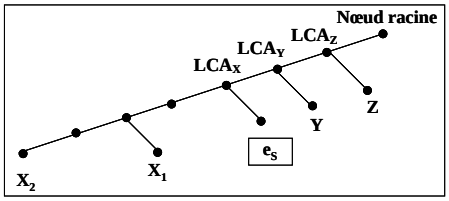

FIG. 5 – *Exemple de taxonomie S_{WN}.*

Du fait du mode de calcul de la mesure, la similarité est plus grande entre e_s et un de ses nœuds frères ou un des descendants proches de ce frère, qu'entre e_s et son grand-père, et ce, jusqu'à une certaine profondeur *p* du descendant, qu'il est possible de déterminer a priori, pour un élément e_s étant donnée sa profondeur. Notre stratégie de parcours de S_{WN} est basée sur cette propriété. Elle consiste tout d'abord à tester si le père P de l'élément e_s dans S_{WN} est un élément de la taxonomie cible, $\mathbf{T_{Cible}}$. Si c'est le cas, il est l'élément le plus proche cherché. Dans le cas contraire, il s'agit de rechercher un élément de $\mathbf{T_{Cible}}$ parmi les descendants du père de e_s jusqu'à la profondeur *p*. Cette profondeur atteinte, si aucun élément de $\mathbf{T_{Cible}}$ n'a été trouvé, il est nécessaire de tester alternativement les descendants de profondeur supérieur à *p* et les ascendants (puis les descendants des ascendants) du père. Cette stratégie permet de limiter le nombre de calculs de similarité à effectuer.

Sur l'exemple FIG. 5, le père de e_s n'appartient pas à $\mathbf{T_{Cible}}$, l'élément e_s est à la profondeur 4, son grand-père à la profondeur 2, la profondeur limite calculée pour e_s (de profondeur 4) est de 5. Parmi les descendants du père de e_s, aucun élément de profondeur inférieure ou égale à 5 n'appartient à $\mathbf{T_{Cible}}$. Après avoir vérifié que le grand-père de e_s n'appartenait pas à $\mathbf{T_{Cible}}$, on étudie les descendants du père de profondeur 6. X_1 appartenant à $\mathbf{T_{Cible}}$, il est donc l'élément recherché.

Cette technique permet d'établir des mises en correspondance entre des éléments connus de Wordnet, des expressions composées en général de peu de mots. Il s'agit de suggestions d'appariement faites à l'expert sans préciser le type de relation exacte reliant les concepts. L'expert peut les accepter en les validant ou les refuser.

3.3 L'exploitation conjointe de la structure des deux taxonomies

A cette étape du processus de découverte des mappings, nous proposons d'appliquer des heuristiques inspirées de celles proposées dans (Melnik et al. (2002), Madhavan et al. (2001), Doan et al. (2002). 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 du concept *Fruit and fruit products* dans la taxonomie source. Sachant que la majorité des descendants du concept *Fruit and fruit products* dans la taxonomie source ont été reliés au concept *Drink* ou à l'une de ses spécialisations dans la taxonomie 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, d'une part, de quel nœud général de la taxonomie cible le concept à apparier est le plus proche, puis s'il

doit être rattaché à ce nœud général (*Drink* dans FIG. 6) ou à un nœud plus spécifique (par exemple *Apple juice* dans FIG. 6).

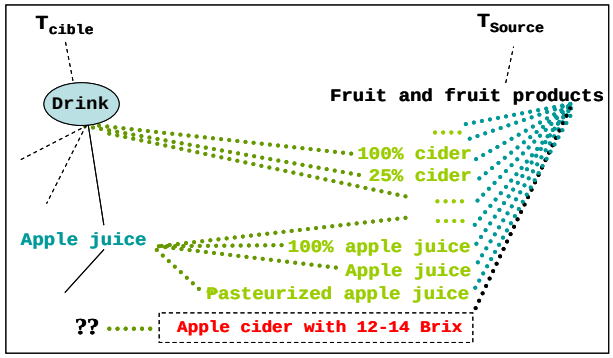

FIG. 6 - *Mappings des frères de Apple cider with 12-14 Brix*

Etant donné un concept e_S feuille de la taxonomie source et ses concepts frères dans cette taxonomie, nous définissons l'ensemble MappingsOfNeighbours (MoN) comme l'ensemble composé des termes de la taxonomie cible auxquels les concepts frères de e_S ont été rattachés par un mapping. Nous mémorisons aussi pour chaque élément de MoN le nombre de mappings établis avec un frère de e_S et nous retenons comme candidats au mapping pour le concept e_S les éléments de MoN intervenant dans au moins deux mappings. Soit CMoN cet ensemble. Nous recherchons ensuite dans la taxonomie cible, les pères des éléments $e_c \in$ CMoN et retenons comme nœuds généraux pertinents, s'ils existent, les nœuds qui sont les pères d'au moins un tiers des éléments de MoN (ou eux-mêmes, par exemple *Drink*). Les éléments de CMoN seront présentés à l'expert regroupés par nœuds généraux pertinents s'ils existent et ordonnés par nombre de mappings établis décroissant.

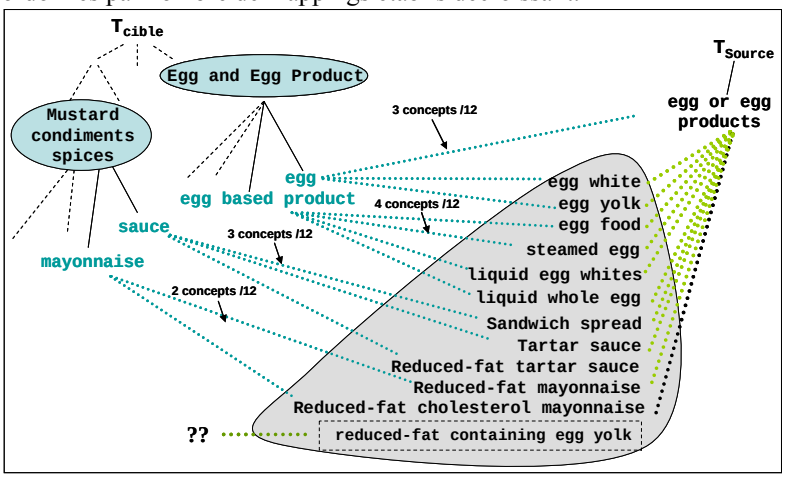

FIG. 7 - *Mappings des noeuds frères de reduced-fat containing egg yolk.*

Sur l'exemple FIG. 7, 3 des 12 frères du concept *reduced-fat containing egg yolk* ont été appariés avec *egg*, 3 avec *egg based product*, 3 avec le concept *sauce* et 2 avec *mayonnaise*. Dans la taxonomie source, *Egg* et *egg based product* ont pour père commun *Egg and Egg*

product. *Sauce* et *mayonnaise* ont pour père *Mustard, condiment, spices*. Le système fait deux propositions de mappings, chacune consistant à suggérer une mise en correspondance soit avec le nœud général (i.e. *Egg and Egg product* ou *Mustard, condiment, spices)*, soit avec l'une de ses deux spécialisations, mais sans toutefois être en mesure de choisir ou de préciser le type de relation exacte reliant le concept.

En revanche, sur l'exemple FIG. 6, sur les 11 frères de *Apple Cider with 12-14 Brix,* 5 ont été mis en correspondance avec *Apple juice*, 3 avec *Drink* et 3 avec *Fruit*. Dans la recherche des nœuds généraux pertinents, le concept *Fresh fruits and vegetables* père de *Fruit* n'est pas retenu car il ne recouvre que 3 sur 11 des éléments adjacents. En revanche, le concept *Drink* est retenu comme car il recouvre 8 sur 11 des éléments de MoN, 3 par lui-même et 5 en tant que père du concept *Apple juice.* Le système ne proposera donc qu'une direction d'appariement, pour le sous-arbre *Drink* et sa spécialisation *Apple juice.*

Si e_S n'est pas un élément feuille de la taxonomie source, la recherche des éléments de MoN est faite sur les fils de e_S, ses frères et leurs descendants. Un mapping sera proposé avec un élément $e_c \in$ CMoN si e_c est le plus petit généralisant commun à l'ensemble des termes de CMoN. Cette technique d'appariement 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.

4 Expérimentations

Un prototype, TaxoMap, a été implémenté en java. Il a permis de réaliser des expérimentations sur les taxonomies Sym'Previus et Com'Base dans le domaine du risque alimentaire[2]. Sym'Previus est l'ontologie cible, Com'Base est l'ontologie source.

La taxonomie de Sym'Previus est composée de 460 termes organisés en 7 niveaux. Com'base est une taxonomie de 172 termes organisée en 2 niveaux, le premier niveau comprenant uniquement 12 termes. Les deux taxonomies ont des parties qui se recouvrent mais ne sont pas structurées de la même façon : une branche peut être détaillée dans l'une et pas dans l'autre ou inversement. La différence de structure s'explique par le fait qu'elles traduisent des points de vue distincts. Enfin, le niveau de granularité n'est pas le même et le nombre de termes représentés est alors différent.

Afin d'évaluer les résultats obtenus en sortie de TaxoMap, un expert en microbiologie a défini manuellement les mappings entre les deux taxonomies, soit pour 172 éléments (nombre d'éléments de Com'Base). L'exécution de Taxomap a permis de générer 96 mappings probables. Ces mappings ont l'avantage d'avoir une précision très importante, supérieure à 90 % (Kefi et al. (2006)) au détriment du rappel (56 %). Les techniques structurelles sont alors très utiles pour compléter les résultats, même si les mappings générés sont moins sûrs. Leur précision, au niveau des expérimentations réalisées, confirme l'ordre d'application proposé dans notre méthode. Les résultats obtenus sont synthétisés dans TAB. 1.

Les mappings non trouvés par la technique exploitant la structure de la taxonomie cible correspondent tous à des mappings générés automatiquement mais qui ne sont pas pertinents. Ils s'expliquent en partie par la difficulté à traiter les labels comportant beaucoup de mots et surtout à distinguer, parmi ces mots, ceux qui font référence au concept sous-jacent et ceux

[2] Ce travail a été réalisé dans le cadre du projet RNTL e.dot (2003-2005).

qui ne font que le caractériser. En effet, cette technique exploite la structure de $\mathbf{T_{Cible}}$, mais elle repose aussi beaucoup sur les calculs de similarité entre éléments, ces calculs étant principalement basés sur des comparaisons de chaînes de caractères. Ce problème de traitement de labels mis à part, cette technique s'est montrée fort utile dans 28 cas sur 42, soit 66 % des cas.

Techniques structurelles	Nombre d'éléments de T_{Source} étudiés	Nombre de mappings suggérés	Nombre de mappings confirmés	Nombre de mappings non trouvés	Précision
Exploitation de la structure de T_{Cible}	42	42	28	14	66 %
Exploitation de la structure de WordNet	14	10	9	5	64 %
Exploitation de la structure de T_S et T_C	5	3	3	2	60 %

TAB. 1 – *Nombre de mappings trouvés par technique employée.*

La technique basée sur WordNet s'est révélée être tout à fait complémentaire des techniques précédemment appliquées. 9 mappings ont pu être trouvés, par exemple *100% cider* a pu être apparié avec *Drink, Cantaloupe* avec *Melon* et *Frankfurter* avec *Sausage*. Les 5 éléments pour lesquels nous n'avons aucun mapping à la fin de l'application de cette technique correspondent soit à des suggestions fausses (*Phosphate buffer* est mis en correspondance avec *Drink* alors que l'expert proposait que ce soit un fils de *Culture Medium*), soit à des éléments pour lesquels aucune suggestion n'a été faite : cas d'acronymes (ex : *TSB*) ou de mots techniques non reconnus par WordNet (ex : *Egyptian Kofta*).
Nous avons appliqué la dernière technique sur les 5 éléments restant à apparier : *Phosphate buffer, Egyptian Kofta, TSB, Tampeh, Pecan nuts*. Deux éléments, *TSB* et *Phosphate buffer*, n'ont pu être mis en correspondance. Ils sont frères l'un de l'autre. Ils ont un seul autre frère commun pour lequel un mapping a été trouvé avec *Culture Medium*. *Culture Medium* serait pertinent (d'après l'expert) pour *Phosphate buffer* et *TSB* mais, en réalité, aucune proposition n'est faite car la technique exige que l'élément proposé au mapping soit mis en correspondance avec au moins deux des frères de l'élément étudié. Pour les trois autres éléments, l'étude des mappings de leurs frères a permis de faire des propositions qui se sont révélées pertinentes. Par exemple, la suggestion est faite d'apparier *Tampeh* avec *Vegetable* (l'expert proposait *Fresh fruit and vegetables*). De même, *Pecan nuts* est apparié avec *Vegetable* (l'expert l'avait apparié avec un de ses fils dans Sym'Previus). Pour *Egyptian Kofta*, le système propose 3 directions d'appariement, l'une avec *Fresh meat*, une seconde avec *Meat-based product* et une troisième avec *Poultry*. La seconde correspond au mapping de l'expert.

5 Travaux proches et discussion

Il existe aujourd'hui de nombreux travaux qui visent à automatiser la génération de mappings. Une synthèse des techniques utilisées est présentée dans Rahm et Bernstein (2001) et Shvaiko et Euzenat (2004). Les techniques sont variées. Nous nous limiterons, dans cette section, aux travaux mettant en oeuvre des techniques structurelles, centrales dans ce papier.

Les techniques structurelles consistent à exploiter la structure des schémas comparés, souvent représentés sous forme de graphes. Les algorithmes mettant en œuvre ces techniques implémentent diverses heuristiques. Une heuristique consiste, par exemple, à considérer que des éléments de deux schémas distincts sont similaires si leurs sous-concepts directs, et/ou leurs sur-concepts directs et/ou leurs concepts frères sont similaires (Do et Rahm (2001), Noy et Musen (2001) (Thanh Le et al. 2004)).. Ces techniques structurelles peuvent être basées sur la notion de point fixe (Melnik et a. (2002)). Dans S-Match (Giunchiglia et Shvaiko (2004)), le problème de matching est vu comme un problème de satisfiabilité d'un ensemble de formules du calcul propositionel. Les graphes et les correspondances à tester sont traduits en formules de la logique propositionnelle en considérant la position des concepts dans le graphe et non seulement leur nom.

Notre travail se distingue des travaux précédemment cités du fait de la dissymétrie dans la structure des taxonomies à rapprocher. La recherche de structures similaires est impossible. Nous proposons donc d'exploiter les données structurelles différemment. La structure de la taxonomie cible, prise isolément, est tout d'abord utilisée pour déterminer le type du concept avec lequel un mapping peut être établi avec un élément e_S de la taxonomie source. Ceci s'effectue en localisant la partie de la taxonomie cible contenant vraisemblablement l'élément avec lequel un mapping pourra être établi. Cette localisation exploite la structure de la taxonomie cible mais s'appuie également sur les calculs de similarité préalablement effectués, basés sur des aspects purement terminologiques. Ne pouvant pas exploiter la structure de la taxonomie source qui, dans notre contexte de travail, est supposée peu structurée, une autre façon d'exploiter la structure est d'avoir recours à des ressources autres que les taxonomies comparées et d'exploiter la structure de ces ressources. C'est ce que nous faisons lorsque nous utilisons WordNet. Enfin, nous proposons une dernière technique basée sur l'exploitation de la structure des deux taxonomies, compte tenu de l'existence d'une dissymétrie. L'idée est d'étudier les mappings préalablement découverts et d'en déduire, compte tenu de la localisation des éléments mis en correspondance dans chacune des taxonomies, des mappings possibles.

6 Conclusion

Ce papier décrit trois techniques structurelles d'alignement de taxonomies supposées dissymétriques du point de vue de leur structure. Le contexte de travail rend impossible la recherche de similarités structurelles. Ainsi, nous proposons d'autres moyens d'exploiter ce type d'information : exploitation de la structure de la taxonomie cible uniquement, pour localiser la partie au sein de laquelle un mapping est possible, exploitation de la structure de Wordnet ou encore exploitation conjointe des informations structurelles des deux taxonomies combinée avec l'étude de mappings préalablement découverts. Ces techniques sont originales dans la mesure où elles se distinguent de la recherche d'une similarité structurelle entre les modèles à aligner. Elles sont applicables pour faire des suggestions de mappings au concepteur. Ces mappings n'ont pas la même vraisemblance que ceux générés par application de techniques terminologiques, c'est pourquoi notre méthode propose d'appliquer les techniques structurelles après les techniques terminologiques. Il s'agit néanmoins d'un bon complément comme les expérimentations l'ont montré.

Références

Doan, A., J. Madhavan, P. Domingos, A. Halevy (2002). *Learnig to map between ontologies on the semantic web*. WWW, pp. 662-673, N.Y., USA, ACM Press.

Do, H. H., E. Rahm (2001). *COMA – A system for flexible combination of schema matching approaches.* VLDB, pp. 610-621.

Giunchiglia, F., P. Shvaiko (2004). *Semantic Matching*. The Knowledge Engineering review, 18(3):265-280.

Kefi, H.,B. Safar, C. Reynaud (2006). *Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes*. RFIA, Tours.

Lin, D (1998). *An Information-Theoretic Definition of Similarity*. ICML, Madison, pp. 296-304.

Madhavan, J., P. A. Bernstein, E. Rahm (2001). *Generic matching with Cupid*. VLDB Journal, pp. 49-58.

Miller, G. A. (1995). *WordNet: A lexical Database for English*. Communications Of the ACM, Vol. 38, n°11, P. 39-45.

Noy, N. F., M. A. Musen (2001). *Anchor-Prompt: Using non-local context for semantic matching*. Workshop on Ontologies and Information Sharing at IJCAI-2001, Seattle, WA.

Melnik, S., H. Garcia-Molina, E. Rahm (2002). *Similarity Flooding*: *A versatile Graph Marching Algorithm and its application to schema matching*. ICDE, San Jose CA.

Rahm, E., P. Bernstein (2001). *A survey of approaches to automatic schema matching*. VLDB Journal: Very Large Data Bases, 10(4):334-350.

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

Thanh Le, B., R. Dieng-Kuntz, F. Gandon (2004). *On Ontology Matching Problems for building a Corporate Semantic Web in a Multi-Communities Organization*. ICEIS (4), pp. 236-243.

Wu, Z., M. Palmer (1994). *Verb semantics and lexical selection.* Computational Linguistics, Las cruces, pp. 133-138.

Summary

This paper deals with generation of mappings in order to align Web taxonomies. The objective is to allow a uniform access to documents in a given application domain. Retrieval of documents is based on taxonomies. We propose to align the taxonomy of a Web portal with the ontology of external documents. That way, the number of accessible documents from the portal can increase without any change in its query interface. This paper presents structural techniques composing the alignment method when dealing with taxonomies which are very different from a structure point of view. A prototype, TaxoMap, has been implemented. It has supported experiments which we present.