Alignement extensionnel et asymétrique de hiérarchies conceptuelles par découverte d'implications entre concepts

Jérôme David*, Fabrice Guillet*
Régis Gras*, Henri Briand*

* LINA CNRS FRE 2729 - Equipe COD
Ecole Polytechnique de l'Université de Nantes
rue Christian Pauc, 44306 NANTES Cedex 3, France
jerome.david,fabrice.guillet,henri.briand@polytech.univ-nantes.fr,
http://www.sciences.univ-nantes.fr/lina/fr/research/teams/EDC/index.html

Résumé. Dans la littérature, de nombreux travaux traitent de méthodes d'alignement d'ontologies. Ils utilisent, pour la plupart, des relations basées sur des mesures de similarité qui ont la particularité d'être symétriques. Cependant, peu de travaux évaluent l'intérêt d'utiliser des mesures d'appariement asymétriques dans le but d'enrichir l'alignement produit. Ainsi, nous proposons dans ce papier une méthode d'alignement extensionnelle et asymétrique basée sur la découverte des implications significatives entre deux ontologies. Notre approche, basée sur le modèle probabiliste d'écart à l'indépendance appelé intensité d'implication, est divisée en deux parties consécutives : (1) l'extraction, à partir du corpus textuel associé à l'ontologie, et l'association des termes aux concepts; (2) la découverte et sélection des implications génératrices les plus significatives entre les concepts. La méthode proposée est évaluée sur deux jeux de données réels portant respectivement sur des profils d'entreprises et sur des catalogues de cours d'universités. Les résultats obtenus montrent que l'on peut trouver des relations pertinentes qui sont ignorées par un alignement basé seulement sur des mesures de similarité.

1 Introduction

Les ontologies ont été créées dans le but de conceptualiser et partager des connaissances de manière structurée (Gruber, 1993). Leur usage en gestion des connaissances s'amplifie avec l'essor du Web sémantique. En effet, les ontologies ont la vertu de se traduire sous des formes très variées depuis de simples taxonomies comme les systèmes catégories (Yahoo, OpenDirectory), en passant par des systèmes de métadonnées interopérables (Dublin Core Metadata initiative) et allant jusqu'aux ontologies lourdes décrivant de véritables théories logiques. Notamment, on trouve des ontologies différentes portant sur le même domaine. Il s'avère donc nécessaire de disposer de techniques pour relier ces ontologies. Dans cette optique, l'alignement vise à trouver des relations entre deux ontologies (entre les classes, les relations, les propriétés...).

Dans la littérature, de nombreux travaux traitent de méthodes d'alignement. Ces approches reposent sur des techniques très différentes (Kalfoglou et Schorlemmer, 2003) comme l'apprentissage bayésien des probabilités jointes entre concepts (Doan et al., 2004), la classification conceptuelle (Stumme et Maedche, 2001), la fusion de schéma de bases de données (Madhavan et al., 2001), les modèles logiques en graphe conceptuels (Fürst et Trichet, 2005), la recherche de morphismes entre graphes représentant les ontologies (Melnik et al., 2002). La distinction entre ces travaux peut être faite au niveau des méthodes d'alignement utilisées pour la comparaison. La classification proposée par (Euzenat et Valtchev, 2003), distingue quatre familles de méthodes : (1) les *méthodes terminologiques* basées sur des mesures de similarité entre chaînes de caractères ou faisant intervenir une ressource terminologique externe ; (2) les *méthodes structurelles* comparant, d'une part, deux concepts à partir de mesures de similarité entre les constituants (attributs, propriétés) des concepts ou à partir de leur position respective dans leur hiérarchie (Noy et Musen, 2000) ; (3) les *méthodes extensionnelles* comparant les concepts à partir de leur ensemble d'instances respectif (Doan et al., 2004) ; (4) les *méthodes sémantiques* basées sur un modèle sémantique théorique utilisé pour la comparaison des concepts (Giunchiglia et al., 2004),(Fürst et Trichet, 2005).

La plupart de ces travaux utilisent des relations symétriques de similarité. Pourtant, d'autres types de relations asymétriques peuvent être utilisées dans le but d'enrichir l'alignement produit. Par exemple, la recherche d'implications (généralisations) permet de trouver les concepts équivalents (exemple : si $auto \rightarrow voiture$ et $voiture \rightarrow auto$ alors $auto \leftrightarrow voiture$), mais elle permet aussi de découvrir si un concept est plus général (ou plus plus spécifique) qu'un autre. Parmi les méthodes prenant en compte la relation d'implication, nous pouvons citer S-MATCH (Giunchiglia et al., 2004). Cette dernière évalue entre autre des relations d'équivalence et d'implication en s'appuyant sur un thésaurus (Wordnet).

L'objectif de notre papier est de proposer une méthode extensionnelle d'alignement basée sur la découverte des relations asymétriques d'implication entre deux ontologies. Nous nous restreignons à des ontologies constituées d'une hiérarchie de concepts dont les concepts sont associés à des documents textuels partageant un vocabulaire commun.

Notre approche est divisée en deux parties consécutives qui utilisent toutes deux le modèle probabiliste d'écart à l'indépendance appelé intensité d'implication (Gras, 1979; Gras et al., 1996) :

- L'extraction des termes pertinents pour chacune des deux hiérarchies. Ce processus permet d'extraire, puis de sélectionner, à partir des documents associés aux hiérarchies, un ensemble de termes pertinents pour chaque concept.
- L'extraction d'implications entre les concepts sous forme de règles d'association (Agrawal et al., 1993). Nous prenons en compte la relation de spécialisation de chaque hiérarchie afin d'extraire les règles les plus générales et ainsi réduire la redondance.

Nous présenterons dans une première section, la méthodologie suivie ainsi que la formalisation des données. Ensuite, nous préciserons les détails de notre approche en commençant par la phase d'analyse et de sélection des termes pertinents, suivie de la découverte de règles entre les deux hiérarchies. Finalement, une évaluation expérimentale nous permettra d'analyser et d'évaluer les performances de notre système.

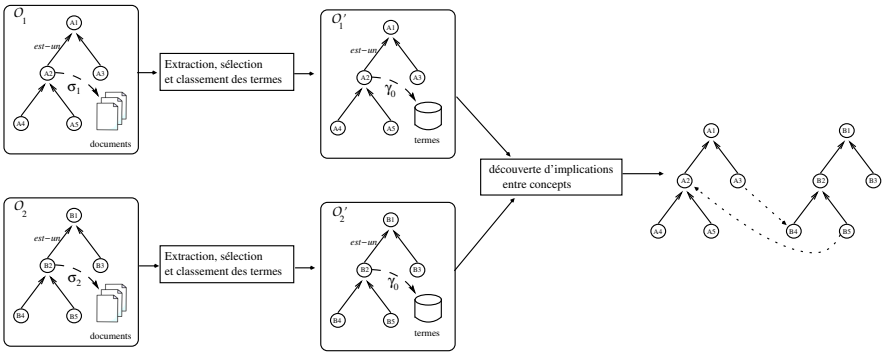

FIG. 1 – *Méthode d'alignement*

2 Méthodologie et formalisation

Nous avons en entrée du processus, deux hiérarchies conceptuelles (figure 1). Chaque hiérarchie est composée d'un ensemble de concepts C structurés par une relation d'ordre partiel $is-a$ (notée par la suite $\leq$). Des parties d'un ensemble de documents D sont associées aux concepts de la hiérarchie. Nous représentons une hiérarchie par un n-uplet $\mathcal{O} = (C, \leq, D, \sigma_0)$, où $\sigma_0 \subseteq C \times D$ est une relation qui associe les concepts aux documents. $\sigma_0(c)$ désigne l'ensemble des documents associés au concept c. Chaque document d est composé d'un ensemble de termes $\{t | t\ apparait\ dans\ d\}$. Nous appellerons T, l'union des ensembles de termes contenus dans les documents et $\delta \subseteq T \times D$, la relation associant les termes aux documents. Nous noterons $\delta(t)$, l'ensemble des documents qui contiennent le terme t. A partir de la relation d'ordre partiel portant sur les concepts, nous définissons la relation :

$$\sigma(c) = \bigcup_{c' \leq c} \sigma_0(c')$$

Une première étape (figure 1) d'extraction, de sélection et d'association des termes aux concepts, nous permet de représenter une hiérarchie $\mathcal{O}$ par le quadruplé $\mathcal{O}' = (C, \leq, T', \gamma_0)$, où $T' \subset T$ représente l'ensemble des termes sélectionnés, et $\gamma_0 \subseteq C \times T'$ est une relation qui associe à chaque concept, ses termes significatifs. Nous notons $\gamma_0(c)$, l'ensemble des termes significatifs du concept c. A partir de la relation d'ordre partiel $\leq$, nous définissons la relation suivante :

$$\gamma(c) = \bigcup_{c' \leq c} \gamma_0(c')$$

La deuxième étape concerne la découverte d'implications significatives entre les concepts issus des hiérarchies $\mathcal{O}'_1 = (C_1, \leq_1, T'_1, \gamma_1)$ et $\mathcal{O}'_2 = (C_2, \leq_2, T'_2, \gamma_2)$. Pour cela, nous nous appuyons sur le modèle des règles d'association (Agrawal et al., 1993). L'extraction des règles est réalisée sur l'ensemble des termes communs aux deux hiérarchies. Ainsi, nous définissons

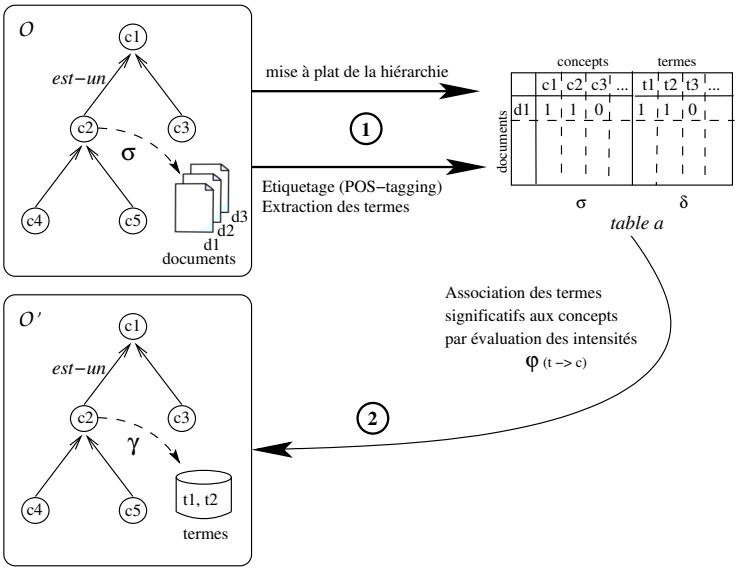

FIG. 2 – *prétraitements linguistiques et extraction des termes pertinents*

$T'_{1\cap 2} = T'_1 \cap T'_2$ et la relation

$$\gamma_{1\cap 2}(c) = \begin{cases} \gamma_1(c) \cap T_2 \; si \; c \in C_1 \\ \gamma_2(c) \cap T_1 \; si \; c \in C_2 \end{cases}$$

Une règle $A \rightarrow B$, où $A \in C_1$ et $B \in C_2$, est une quasi-implication de l'ensemble $\gamma_{1\cap 2}(A)$ dans l'ensemble $\gamma_{1\cap 2}(B)$. La règle $A \rightarrow B$ signifie que les termes significatifs du concept A ont tendance à être significatifs du concept B.

3 Extraction et sélection des termes significatifs

Notre objectif consiste à extraire un ensemble de termes significatifs pour chacun des concepts. L'idée principale de ce processus est la suivante : Un terme t sera significatif d'un concept c si il existe relativement peu de documents contenant le terme t qui ne sont pas associés au concept c. Nous choisissons d'associer le terme t au concept c si la règle d'association $t \rightarrow c$ est significative selon l'intensité d'implication.

Afin d'extraire les règles $t \rightarrow c$, nous traduisons les relations σ et δ par la *table a*, figure 2. Ainsi, chaque document est représenté avec d'une part les concepts auxquels il est associé, et d'autre part les termes qu'il contient. L'ensemble des termes T_0 est constitué de verbes et de termes binaires (termes composés de deux mots significatifs). L'extraction de ces termes binaires se justifie par le fait qu'ils sont plus porteurs d'information et donc moins ambigus que de simples mots. L'acquisition des termes binaires est réalisée par le logiciel ACABIT

(Daille, 2003) à partir des textes préalablement étiquetés (grammaticalement) et lemmatisés par la suite logicielle MontyLingua (Liu, 2004).

La deuxième étape (étape 2, figure 2) consiste à évaluer toutes les règles d'association binaires $t \rightarrow c$, avec $t \in T_0$ et $c \in C$, afin de constituer pour chaque concept c, un ensemble de termes significatifs $\gamma_0(c)$ défini par :

$$\gamma_0(c) = \{t \in T_0 | \varphi(t \rightarrow c) > \varphi_t\}.$$

où φ_t est la valeur seuil et $\varphi(t \rightarrow c)$ est la valeur d'intensité d'implication définie par :

$$\varphi(t \rightarrow c) = 1 - Pr\left(N_{t \wedge \overline{c}} \leq n_{t \wedge \overline{c}}\right)$$

où $n_{t \wedge \overline{c}} = card(I(t) - \sigma(c))$ représente le nombre observé de documents contenant le terme t qui ne sont pas associés au concept c et $N_{t \wedge \overline{c}}$ représente le nombre attendu (sous hypothèse d'indépendance des descriptions t et c) de documents contenant le terme t qui ne sont pas associés au c. Comme les phénomènes étudiés sont rares, nous modélisons la variable aléatoire $N_{t \wedge \overline{c}}$ par une loi de Poisson de paramètre $\lambda = n_t.n_{\overline{c}}/n$ où n_t est le nombre de documents contenant le terme t, $n_{\overline{c}}$, le nombre de documents non associés au concept c et n le nombre total de documents.

Notons que les fréquences des termes dans les documents ne sont pas prises en compte, nous nous intéressons seulement à la présence ou absence des termes dans les documents.

4 Découverte d'implications significatives entre concepts

4.1 Critères de sélection d'une implication significative

A partir de deux hiérarchies $\mathcal{O}'_1$ et $\mathcal{O}'_2$, cette deuxième partie a pour objectif de découvrir des implications sous forme de règles binaires entre les concepts des deux structures. Pour cela, nous nous inspirons de la méthode de découverte de règles d'association généralisées proposée par (Srikant et Agrawal, 1995).

Nous définissons pour $A \in C_1$ et $B \in C_2$, l'intensité d'implication de la règle $A \rightarrow B$ par :

$$\varphi(A \rightarrow B) = 1 - Pr\left(N_{A \wedge \overline{B}} \leq n_{A \wedge \overline{B}}\right)$$

où $n_{A \wedge \overline{B}} = card(\gamma_{1 \cap 2}(A) - \gamma_{1 \cap 2}(B))$ représente le nombre observé de termes associés au concept A qui ne sont pas associés au concept B. $N_{A \wedge \overline{B}}$ représente le nombre attendu (sous hypothèse d'indépendance des descriptions A et B) de termes associés à A qui ne sont pas associés à B. Nous modélisons la variable aléatoire $N_{A \wedge \overline{B}}$ par une loi de Poisson de paramètre λ

$$\lambda = card(\gamma_{1 \cap 2}(A)).card(T'_{1 \cap 2} - \gamma_{1 \cap 2}(B))/card(T'_{1 \cap 2})$$

Intéressons nous, par exemple, à l'implication $A2 \rightarrow B4$ de la figure 3. Nous avons $n_{A2 \wedge \overline{B4}} = 1$ contre-exemple et $\lambda = 6,6$. La valeur d'intensité d'implication est calculée de la manière suivante :

$$\varphi(A2 \rightarrow B4) = \sum_{k=max(0, n_a - n_b)}^{n_{a \wedge \overline{b}}} e^{-\lambda}.\frac{\lambda^k}{k!} = 0,97$$

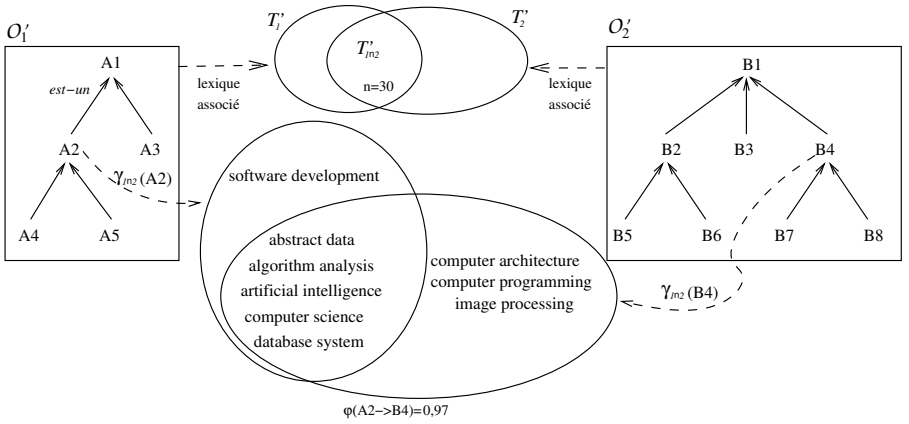

FIG. 3 – *Implication entre concepts*

La significativité d'une règle est exprimée par la valeur d'intensité d'implication et par le caractère spécifique de sa conclusion combiné à la généralité de sa prémisse. Ainsi pour $A \in C_1$ et $B \in C_2$, une règle $A \rightarrow B$ sera dite "significative" si :

1. $\varphi(A \rightarrow B) > \varphi_r$
2. $\forall X, \forall Y$ avec $A \leq_1 X$ et $Y \leq_2 B$, $\varphi(X \rightarrow Y) \leq \varphi(A \rightarrow B)$

Le deuxième critère traduit la capacité d'une règle à générer d'autres règles. Il permet ainsi de réduire la redondance dans l'ensemble des règles extraites. En effet, à partir de la règle $A \rightarrow B$, nous pouvons déduire toutes les règles de la forme : $X \rightarrow Y$ avec $B \leq Y$ et $X \leq A$ car $\gamma_{1\cap2}(B) \subseteq \gamma_{1\cap2}(Y)$ et $\gamma_{1\cap2}(X) \subseteq \gamma_{1\cap2}(A)$. Ainsi, nous dirons que la règle $A \rightarrow B$ est génératrice de l'ensemble des règles de type $X \rightarrow Y$. Par exemple, sur la figure 3, la règle $A2 \rightarrow B4$ permet de générer les règles $A2 \rightarrow B1$, $A4 \rightarrow B4$, $A5 \rightarrow B4$, $A4 \rightarrow B1$ et $A5 \rightarrow B1$.

Contrairement à (Srikant et Agrawal, 1995) qui s'appuie sur les mesures de support et de confiance, nous préférons utiliser la mesure d'intensité d'implication φ. La confiance et le support ne sont pas adaptés pour découvrir des règles génératrices car ils favorisent les règles ayant des conclusions générales. En effet, la confiance d'une règle $A \rightarrow B$ est plus grande ou égale à la confiance d'une règle de la forme $A \rightarrow Z$ avec $Z \leq_2 B$. Par contre, φ prend en compte la taille de la conclusion et tend vers 0 quand la conclusion devient trop générale : par exemple, sur la figure 3, $\forall A \in C_1$, $\varphi(A \rightarrow B1) = 0$ car $\gamma_{1\cap2}(B1) = T'_{1\cap2}$.

4.2 Algorithmes

La prise en compte de la relation d'ordre partiel entre les concepts des hiérarchies permet d'optimiser la phase de sélection des implications significatives. En effet, une recherche descendante de règles (du haut de la hiérarchie vers le bas) nous permet d'éviter l'évaluation

Variables d'entrée :
A : un concept de $\mathcal{O}_1$.
$\mathcal{B}_{courant}$: un ensemble de concepts de $\mathcal{O}_2$.
Procédure specializeAntecedent(A, B)
Début
Pour chaque $B_x \in \mathcal{B}_{courant}$ **faire**
$specializeConsequent(A, B_x, \mathcal{B}_{courant}, 0.0)$
FinPour
Pour chaque $fils \in LesFils(A)$ **faire**
$\mathcal{B}'_{courant} := \mathcal{B}_{courant}$
$specializeAntecedent(fils, \mathcal{B}'_{courant})$
FinPour
Fin

FIG. 4 – *Algorithme de spécialisation de la prémisse*

des règles ayant une conclusion trop spécifique. Nous détaillons dans cette section la méthode de recherche des implications significatives entre concepts issus d'une hiérarchie $\mathcal{O}_1$ vers les concepts d'une hiérarchie $\mathcal{O}_2$. Notre méthode est divisée en deux algorithmes.

Le premier algorithme (figure 4) prend en entrée un concept A de la hiérarchie $\mathcal{O}_1$, et un ensemble de concepts $\mathcal{B}_{courant} \subset C_2$ de $\mathcal{O}_2$. Pour chacun des concepts de $\mathcal{B}_{courant}$, la recherche et la sélection des conclusions valides sont effectuées par le second algorithme (figure 5). Ensuite, la procédure est relancée récursivement sur les fils de A et une copie de l'ensemble $\mathcal{B}_{courant}$ mis à jour par le second algorithme. L'ensemble de concepts $\mathcal{B}_{courant}$ recense les sous-parties de la hiérarchie $\mathcal{O}_2$ qui ne contiennent aucun concept figurant dans les conclusions des règles sélectionnées lors des récursions précédentes. Cette liste permet d'éviter d'évaluer des règles ne satisfaisant pas le critère 2 explicité dans la section 4.1.

Le deuxième algorithme (figure 5) prend en entrée un concept A de la hiérarchie $\mathcal{O}_1$, et un concept B de $\mathcal{O}_2$. Il recherche parmi l'ensemble $\{B_x | B_x \leq_2 B\}$, un sous-ensemble de conclusions valides pour A. Une conclusion B_s sera sélectionnée, si elle respecte les deux critères de la section 4.1.

La recherche se faisant de manière descendante dans la hiérarchie des descendants de B, nous choisissons d'arrêter la recherche dans une des branches si $\varphi(A \rightarrow B_x)$, est en dessous de la valeur seuil φ_r et si aucune spécialisation de la conclusion ne permettra de repasser au dessus du seuil φ_r. Pour cela, nous nous appuyons sur une propriété de l'intensité d'implication qui définit $A \cup B_x$ comme étant la meilleure spécialisation de la conclusion de $A \rightarrow B_x$.

Lors de la recherche de règles, nous ignorons les racines des hiérarchies. En effet, les concepts racines des hiérarchies sont associés à l'ensemble des termes étudiés. Ainsi, la valeur d'intensité d'implication, ne peut être évaluée ou est nulle si un concept racine est en prémisse ou en conclusion.

Variable globale :
φ_r : La valeur seuil d'intensité d'implication
Variables d'entrée :
A : un concept de $\mathcal{O}_1$.
B : un concept de $\mathcal{O}_2$.
φ_{max} : le meilleur score $\phi(A \rightarrow B_p)$ avec $B \leq B_p$
Variables d' entrée/sortie :
$\mathcal{B}_{courant}$: La liste de concepts "courants" de la deuxième hiérarchie.
$listeImplications$: la liste des règles sélectionnées.
Valeur de retour :
La valeur φ de la meilleure règle $A \rightarrow B_x$ avec $B_x \leq B$

Fonction specializeConsequent(A, B, $\mathcal{B}_{courant}$,φ_{max}) **Début**
$meilleurFils := FAUX$
$\varphi_{courant} := \varphi(A, B)$
$returnVal := \varphi_{courant}$
Si $(\varphi_{courant} < \varphi_r)$ **alors**
$\varphi' := \varphi(A, A \cap B)$
Si $(\varphi' < \varphi_r)$ **alors**
retourne $\varphi_{courant}$
FinSi
FinSi
Pour chaque $fils \in LesFils(B)$ **faire**
$\varphi_{fils} := specializeConsequent(A, fils, \mathcal{B}_{courant})$
Si $(\varphi_{fils} \geq \varphi_{courant})$ **alors**
$meilleurFils := VRAI$
$\mathcal{B}_{courant} := \mathcal{B}_{courant} - \{B\}$
Si $(retunVal < \varphi_{fils})$ **alors**
$returnVal := \varphi fils$
FinSi
FinSi
Si $(\varphi_{courant} > \varphi_r)$ et $\neg(meilleurFils)$ et $(\varphi_{courant} \geq \varphi_{max})$ **alors**
$listeImplications := listeImplications \cup (A \rightarrow B)$
$\mathcal{B}_{courant} := \mathcal{B}_{courant} \cup \{B\}$
$\varphi_{max} := \varphi_{courant}$
FinSi
FinPour
retourne $returnVal$
Fin

FIG. 5 – *Algorithme de spécialisation de la conclusion*

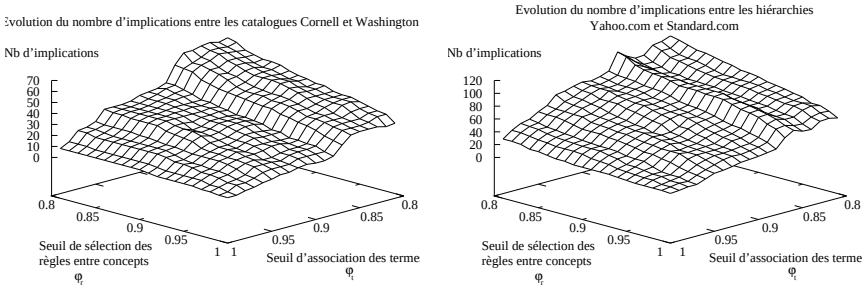

FIG. 6 – *Influence des valeurs seuil sur le nombre d'implications retenues*

5 Évaluation expérimentale

5.1 Les données analysées

Nous avons testé notre méthode sur deux jeux de test proposés dans (Doan et al., 2004). Le premier jeu de test "Course Catalog" décrit les cours proposés par l'université de Cornell et l'université de Washington. Les hiérarchies contiennent respectivement 166 et 176 concepts ainsi que 4360 et 6957 instances sous forme de documents textuels décrivant des cours. Les cours (instances) sont organisés en écoles et collèges, ensuite en départements et centres à l'intérieur de chaque collège. Le deuxième jeu de test "Company Profile" est issu des annuaires web Yahoo.com et Standard.com. Ces dernières hiérarchies contiennent respectivement 115 et 333 noeuds ainsi que 13634 et 9504 instances. On peut remarquer que la hiérarchie Standard.com a un découpage plus fin que celle de Yahoo.com. Les instances décrivent les activités d'entreprises. Les descriptions d'entreprises sont organisées en secteurs puis en industries.

5.2 Résultats et analyse

Afin d'analyser le comportement de notre méthode de manière quantitative et qualitative, nous avons étudié dans un premier temps, l'influence du choix des valeurs seuils d'intensité d'implication φ_t et φ_r sur le nombre de règles sélectionnées. Pour les deux jeux de test présentés ci-dessus, nous avons testé notre algorithme en faisant varier (de 0,8 à 1) les valeurs seuils pour la sélection des groupes de termes pertinents et pour l'appariement de concepts.

Sur les deux graphiques de la figure 6, on peut remarquer que le choix du seuil φ_t influence plus la quantité d'implications extraites. Prenons l'exemple de l'alignement des catalogues Cornell et Washington : pour une augmentation du seuil φ_t (resp. φ_r) de 0,1 unité, le nombre d'implications extraite baisse en moyenne de 2,15 (resp. 1,2).

Dans un deuxième temps nous avons réalisé un test qualitatif à partir du jeu de test "Course catalog" et des alignements manuels fournis sur le site de A. Doan. Nous avons confronté les résultats produits par notre approche aux alignements manuels. Comme ces derniers sont de nature symétriques, nous avons procédé aux recherches d'implications dans les deux sens (Cornell vers Washington, puis Washington vers Cornell). Ensuite, à partir des implications extraites, nous déduisons des relations d'équivalences.

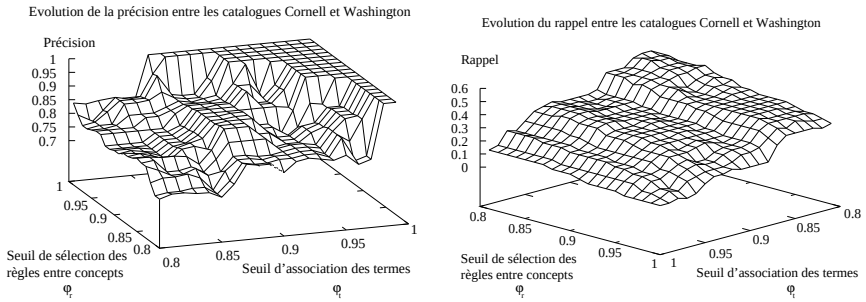

FIG. 7 – *Influence des valeurs seuil sur la précision et le rappel*

Sur la figure 7, nous notons, comme pour le nombre d'implications extraites (figure 6), une plus grande influence du seuil φ_t par rapport au seuil φ_r. Nous pouvons observer une évolution de la précision sur des bons scores (de 0,71 à 1). Le choix des deux seuils influencent quasiment de la même manière l'évolution de la précision. Cependant, le rappel présente une faible moyenne de 0,29 (meilleur score égale à 0,54).

Ces scores mitigés au niveau du rappel sont justifiés, premièrement, par l'existence de concepts pour lesquels il n'y a aucun terme associé. Ce problème est dû à un manque de spécificité du vocabulaire utilisé dans les descriptions des concepts et à une sélection trop stricte des règles $terme \rightarrow concept$. Le deuxième point explicatif est que nous avons considéré, pour le test, seulement les relations d'équivalence. En effet, pour un seuil d'association des termes aux concepts égal à 0,83 et un seuil de sélection des implications entre concepts égal à 0,82, nous obtenons 24 faux positifs. Parmi ces faux positifs, 6 ont été considérés comme des implications et non des équivalences. Notre méthode permet de distinguer, contrairement aux alignements fournis, différents types de relations.

Nous avons aussi extrait des implications significatives qui ne figurent pas dans les alignement manuels. Nous avons par exemple obtenu :
De Cornell à Washington :

City and Regional Planning -> Urban Planning URBDP
Cognitive Studies Program -> Psychology PSYCH
Department of Aerospace Studies -> Aerospace Studies A S
Electrical and Computer Engineering -> Electrical Engineering E E

De Washington à Cornell :

Atmospheric Sciences ATM S -> Earth and Atmospheric Sciences

Ces implications entre concepts sont intéressantes même si elles ne figurent pas dans les alignements manuels. Nous pouvons aussi noter que notre méthode n'est pas sensible aux noms des concepts (une approche ne faisant intervenir qu'une similarité entre chaînes de caractères ne pourrait pas trouver la règle "Cognitive Studies Program -> Psychology PSYCH"). Ainsi, notre méthode prend en compte la sémantique des concepts.

6 Conclusion

Nous avons proposé dans ce papier une approche d'alignement d'ontologies basée sur la découverte de relations d'implication significatives entre concepts provenant de deux ontologies distinctes. Notre méthode, est décomposée en deux phases : (1) L'extraction à partir des documents puis l'association d'un ensemble de termes pertinents pour chaque concept. (2) La découverte de règles d'implication significatives entre concepts en nous appuyant sur les termes sous-jacents. Les intérêts de notre méthode sont d'une part la prise en compte de la sémantique en utilisant des termes binaires contenus dans les documents associés aux concepts, et d'autre part la recherche d'implications permettant d'enrichir l'alignement produit. Notre démarche a été testée sur deux jeux de données réels portant respectivement sur des profils d'entreprises et sur des catalogues de cours d'universités. La confrontation des résultats obtenus et des alignements manuels fournis avec le jeu de test montrent tout d'abord que notre approche distingue parmi les alignements manuels des relations d'équivalence et des relations d'implication. De plus notre méthode permet d'extraire des relations pertinentes qui sont ignorées par l'ensemble d'alignements manuels.

Actuellement, nous n'avons considéré que des hiérarchies conceptuelles associées à un corpus de documents. Nous projetons d'étendre notre approche afin d'exploiter la définition et la structure interne des concepts, puis de permettre la recherche d'implications entre les relations.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216. ACM Press.

Daille, B. (2003). Conceptual structuring through term variations. In F. Bond, A. Korhonen, D. MacCarthy, et A. Villacicencio (Eds.), *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pp. 9–16.

Doan, A., J. Madhavan, P. Domingos, et A. Halevy (2004). *Ontology Matching : a machine learning approach*, pp. 397–416. Springer-Velag.

Euzenat, J. et P. Valtchev (2003). An integrative proximity measure for ontology alignment. In *Semantic Integration Workshop, Second International Semantic Web Conference (ISWC-03)*.

Fürst, F. et F. Trichet (2005). Aligner des ontologies lourdes : une méthode basée sur les axiomes. In *16èmes journées francophones d'ingénierie des connaisssances*, pp. 121–132.

Giunchiglia, F., P. Shvaiko, et M. Yatskevich (2004). S-match : an algorithm and an implementation of semantic matching. In *European Semantic Web Symposium*, LNCS 3053, pp. 61–75.

Gras, R. (1979). Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques mathématiques. Thèse d'Etat, Université de Rennes.

Gras et al., R. (1996). *L'implication statistique, une nouvelle méthode exploratoire de données*. La pensée sauvage.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5*(2), 199–220.

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

Liu, H. (2004). Montylingua : An end-to-end natural language processor with common sense.

Madhavan, J., P. A. Bernstein, et E. Rahm (2001). Generic schema matching with cupid. In *The VLDB Journal*, pp. 49–58.

Melnik, S., H. Garcia-Molina, et E. Rahm (2002). Similarity flooding : A versatile graph matching algorithm and its application to schema matching. In *the 18th International Conference on Data Engineering (ICDE'02)*, pp. 117–128. IEEE Computer Society.

Noy, N. F. et M. A. Musen (2000). Prompt : Algorithm and tool for automated ontology merging and alignment. In *the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pp. 450–455. AAAI Press / The MIT Press.

Srikant, R. et R. Agrawal (1995). Mining generalized association rules. In *the 21th International Conference on Very Large Data Bases (VLDB '95)*, San Francisco, CA, USA, pp. 407–419. Morgan Kaufmann Publishers Inc.

Stumme, G. et A. Maedche (2001). FCA-MERGE : Bottom-up merging of ontologies. In *IJCAI*, pp. 225–234.

Summary

A lot of works deal with ontology alignment. They mostly consider similarity relations which are symmetric. Too few works assess the interest of using asymmetric measures in order to enhance the alignment. In this paper, we suggest an extensional and asymmetric alignment method based on the discovery of significant implications between concepts. Our approach uses a probabilistic model of deviation from independence named implication intensity. Our method is divided into two main parts: (1) the extraction from the textual corpus associated with the ontology and the selection of meaningful terms for each concept. (2) the discovery of significant implications between the concepts. Our method is tested on two benchmarks. The results show that some relevant relations, ignored by a similarity-based alignment, can be found thanks to our approach.