Fouille de méta-données pour la découverte de mappings entre taxonomies : une approche combinant logique et probabilités

Rémi Tournaire[1,2], Marie-Christine Rousset[1]

[1] Laboratoire d'Informatique de Grenoble UMR5217, équipe HADAS
Bâtiment IMAG D - 681, rue de la Passerelle - 38400 Saint Martin d'Hères

[2] LIRIS UMR5205, équipe Base de données,
Bât. Blaise Pascal - INSA de Lyon - 69621 Villeurbanne Cedex

Résumé : Dans le cadre de la découverte automatique de mappings entre ontologies dans un réseau P2P, nous fournissons un formalisme et une méthode d'estimation pour la probabilité d'un mapping, grâce à une démarche bayésienne exploitant les méta-données associées aux ressources des classes les annotant. Nous présentons un algorithme pour fournir de façon efficace, à partir d'un ensemble de mappings candidats, l'ensemble des mappings dont la probabilité dépasse un certain seuil grâce à un ordre sur les mappings fondé sur leur sémantique logique.
Mots-clés : P2P, mappings, ontologies, probabilité, méta-données

1 Contexte et position du problème

Cet article se situe dans le contexte de réseaux pair à pair de partage sémantique de ressources (documents, données, services). Dans les réseaux pair à pair que nous considérons dont SomeWhere (Adjiman *et al.* (2005)) est un exemple, chaque pair annote sémantiquement les ressources qu'il stocke localement par les noms de classes d'une taxonomie qui lui est propre. Des mappings sont des correspondances entre classes de taxonomies de plusieurs pairs qui permettent des reformulations de requêtes dans les vocabulaires adéquats afin de trouver dans l'ensemble du réseau les ressources satisfaisant une requête exprimée en fonction du vocabulaire d'un pair particulier. De nombreux travaux ont porté sur la découverte automatique de mappings entre deux ontologies (e.g. Shvaiko & Euzenat (2005); Doan *et al.* (2002)). La plupart des méthodes d'alignement sont semi-automatiques : elles renvoient comme résultat un ensemble de mappings probables qu'il faut ensuite valider manuellement.

Dans cet article nous établissons un cadre formel pour le calcul de la probabilité d'un mapping à partir des observations sur les méta-données associées aux ressources des pairs concernés. Nous prouvons sa monotonie par rapport à la relation d'implication logique entre mappings. Puis nous présentons la méthode qui en découle pour explorer

de façon efficace un ensemble de mappings candidats dont on veut sélectionner ceux dont la probabilité dépasse un certain seuil.

1.1 Préliminaires

Une **ressource** est un document identifié par une URI[1] qui est unique.
Nous supposons qu'il est possible d'extraire des **méta-données** de chacun des documents. C'est le cas de fichiers de musique en MP3 pour lesquels la norme ID3 propose un langage attribut-valeur de méta-données qui contient de l'ordre de 50 noms d'attributs (e.g., "title", "genre", "artist", "year", ...) dont les valeurs associées sont soit énumérées (comme les valeurs de l'attribut "genre"), soit du texte libre (comme les valeurs des attributs "title" ou artiste"), soit numériques comme les valeurs de l'attribut "year".
L'exemple suivant illustre un extrait de méta-données que l'on peut trouver dans deux fichiers de musique MP3 identifiés respectivement par id_1 et id_2.

	title	artist	genre	year
$md(id_1)$	"It's raining again"	"Supertramp"	"Rock"	"1982"
$md(id_2)$	"Le lundi au soleil"	"Claude François"	"Pop"	"1972"

Les méta-données que nous considérons dans notre cadre formel sont exprimées dans un langage propositionnel relativement à m attributs booléens $A_1, \ldots, A_m$: chaque ressource id_i disponible dans le réseau est associée à un vecteur booléen de méta-données $mdb(id_i)$= $[b_{i_1}, \ldots, b_{i_m}]$ de taille m où $b_{i_j} = mdb(id_i)[j] = 1$ ssi l'attribut A_j est présent dans la description booléenne des méta-données associées au document d'identifiant id_i.
Ces méta-données peuvent être obtenues par pré-traitement et encodage des méta-données réelles que l'on extrait des fichiers. Ainsi, les méta-données exprimées dans la norme ID3 dans l'exemple précédent peuvent être codées de façon booléenne de la façon suivante :

- Attributs booléens : $A_1 = Rock\_genre$, $A_2 = Pop\_genre$,
  $A_3 = Supertramp\_artist$, $A_4 = ClaudeFrancois\_artist$,
  $A_5 = RainingAgain\_title$, $A_6 = LundiSoleil\_title$, $A_7 = 1982\_year$,
  $A_8 = 1972\_year$,
- Méta-données booléennes :
  - $mdb(id_1) = [1, 0, 1, 0, 1, 0, 1, 0]$
  - $mdb(id_2) = [0, 1, 0, 1, 0, 1, 0, 1]$

Le prétraitement et l'encodage propositionnel sont des problèmes importants que nous n'aborderons pas dans cet article.
Chaque **pair** P_i est une entité (logicielle ou matérielle) autonome qui stocke localement des ressources et les catégorise à l'aide d'un ensemble de classes définies et structurées dans son ontologie.

[1] Unified Resource Identifier

L'**ontologie** O_i d'un pair P_i est un ensemble d'axiomes d'inclusions entre classes ou complémentaires de classes : $E_i^k \sqsubseteq E_i^l$ (où E_i^k et E_i^l sont les noms de classes du vocabulaire de P_i ou des complémentaires de noms de classes de ce même pair).
Par exemple l'axiome $Classique_1 \sqsubseteq Musique_1$ exprime que dans l'ontologie O_1 du pair P_1, la classe appelée $Classique$ est une sous-classe de la classe de nom $Musique$, alors que l'axiome $Classique_1 \sqsubseteq \overline{Rock_1}$ exprime que dans O_1, les classes $Classique$ et $Rock$ sont disjointes.
On fait l'hypothèse que les vocabulaires (i.e. les noms de classe) de pairs différents sont distincts. On convient de la notation C_i pour dénoter la classe appelée C par le pair P_i.
Un **mapping** est une *expression d'inclusion* entre classes ou complémentaires de classes de pairs différents. Par exemple, $Mouv_2 \sqsubseteq Rock_1$ exprime que la classe appelée $Mouv$ par le pair P_2 est une sous-classe de la classe appelée $Rock$ par le pair P_1.
Chaque ressource est stockée dans un (ou plusieurs) pair(s) qui déclare(nt) des **axiomes d'appartenance à une classe** de la forme $C_i(id)$ pour exprimer que la ressource d'identifiant id est classée dans C_i. Par exemple, si id_1 et id_2 sont les identifiants des fichiers $SupertrampItsraining.mp3$ et $lelundiausoleil.mp3$ respectivement, les déclarations $Rock_1(id_1)$, $Pop\_Rock_3(id_1)$, $Nostalgie_2(id_2)$ expriment que le premier fichier est catégorisé par le pair P_1 comme du $Rock$ et par le pair P_3 comme du Pop_Rock, alors que le second fichier est catalogué par le pair P_2 dans la classe $Nostalgie$.
Les axiomes d'inclusion, les axiomes d'appartenance ainsi que les mappings ont une **sémantique logique standard** à base d'interprétation ensemblistes des classes. Une interprétation $\mathcal{I}$ est un couple $(\Delta^{\mathcal{I}}, .^{\mathcal{I}})$, où $\Delta^{\mathcal{I}}$ est le domaine d'interprétation et $.^{\mathcal{I}}$ est une fonction qui interprète chaque nom de classe comme un sous-ensemble de $\Delta^{\mathcal{I}}$, et chaque URI comme un élément de $\Delta^{\mathcal{I}}$. L'hypothèse d'URI unique pour une ressource se traduit formellement par : $a \neq b \;\Rightarrow\; a^{\mathcal{I}} \neq b^{\mathcal{I}}$.
L'interprétation du complémentaire d'une classe est l'ensemble complémentaire de son interprétation dans $\Delta^{\mathcal{I}}$.
Etant donnée $\mathcal{O}$ une ontologie (ou une union d'ontologies et de mappings) et $\mathcal{F}$ un ensemble d'axiomes d'appartenance, $\mathcal{I}$ est un modèle de $\mathcal{O} \cup \mathcal{F}$ si :
– pour tout axiome d'inclusion $E \sqsubseteq F$ de $\mathcal{O}$: $E^{\mathcal{I}} \subseteq F^{\mathcal{I}}$
– pour tout axiome d'appartenance $C(a)$ de $\mathcal{F}$: $a^{\mathcal{I}} \in C^{\mathcal{I}}$
Une expression d'inclusion $G \sqsubseteq H$ est vraie dans $\mathcal{O}$ ssi dans tout modèle $\mathcal{I}$ de $\mathcal{O}$, $G^{\mathcal{I}} \subseteq H^{\mathcal{I}}$. On note alors : $\mathcal{O} \models G \sqsubseteq H$.
Une expression d'appartenance $D(b)$ est vraie dans $\mathcal{O} \cup \mathcal{F}$ ssi dans tout modèle $\mathcal{I}$ de $\mathcal{O} \cup \mathcal{F}$, $b^{\mathcal{I}} \in D^{\mathcal{I}}$. On note alors : $\mathcal{O} \cup \mathcal{F} \models D(b)$.
Soit $\mathcal{P}$ un pair ou un ensemble de pairs, $\mathcal{O}(\mathcal{P})$ l'ontologie (ou l'union d'ontologies et de mappings) associée, et $\mathcal{F}(\mathcal{P})$ l'ensemble des axiomes d'appartenance déclarés dans $\mathcal{P}$. Soit $\mathcal{D}(\mathcal{P})$ l'ensemble des identifiants de ressources stockées dans $\mathcal{P}$.
L'**extension** d'une classe C dans $\mathcal{P}$ est l'ensemble des identifiants dont on peut déduire l'appartenance à cette classe :

$$ext(C, \mathcal{P}) = \{d \in \mathcal{D}(\mathcal{P}) \;\; / \;\; \mathcal{O}(\mathcal{P}) \cup \mathcal{F}(\mathcal{P}) \models C(d)\}$$

L'extension du complémentaire de $\mathcal{C}$ dans $\mathcal{P}$ est l'ensemble des identifiants de $\mathcal{P}$ dont on ne peut pas déduire l'appartenance à $\mathcal{C}$: $ext(\overline{C}, \mathcal{P}) = \mathcal{D}(\mathcal{P}) \setminus ext(C, \mathcal{P})$.

Les identifiants appartenant à l'extension d'une classe dans $\mathcal{P}$ sont appelées ses **instances**.

1.2 Position du problème

Dans notre cadre, les relations d'inclusion entre classes d'une ontologie ainsi que les relations d'appartenance d'instances à une classe sont posées comme des axiomes ou déduites logiquement, et sont donc vraies ou fausses. Pour les mappings, nous voulons pouvoir modéliser que certains, bien que ne se déduisant pas logiquement d'autres mappings posés comme axiomes, ont une forte vraisemblance en fonction des méta-données associées aux instances stockées dans les pairs mis en correspondance par ces mappings. Par exemple, l'observation que les caractéristiques communes des méta-données extraites des fichiers MP3 classés par P_2 dans $Nostalgie_2$ se retrouvent pour la plupart dans les méta-données extraites des fichiers mp3 classés par P_1 dans Pop_1 (par exemple, 95% et 99% des instances de ces deux classes ont le genre Pop) est un indice fort en faveur de l'ajout du mapping $Nostalgie_2 \sqsubseteq Pop_1$ comme un nouvel axiome.
Nous explicitons dans la section 2 la sémantique probabiliste que nous définissons pour les mappings et son lien avec leur sémantique logique posée dans la section 1.1. Puis, nous montrons comment calculer la probabilité d'un mapping par une estimation bayésienne en considérons les méta-données associées aux classes des pairs impliqués dans le mapping comme des observations.
Le problème central considéré dans cet article est de déterminer parmi un ensemble de mappings candidats ceux dont la probabilité dépasse un certain seuil, en faisant le moins de calcul de probabilités possibles. Nous exhibons dans la section 3 un algorithme d'énumération et de test de mappings candidats qui exploite la propriété de monotonie de la fonction de probabilité de mappings par rapport à l'implication logique.

2 Modélisation et calcul de la probabilité d'un mapping

Considérons un mapping $E_1 \sqsubseteq F_2$ entre deux pairs P_1 et P_2, où E_1 est une classe (ou le complémentaire d'une classe) de l'ontologie O_1 de P_1, et F_2 est une classe (ou le complémentaire d'une classe) de l'ontologie O_2 de P_2.
A partir des méta-données booléennes associées aux instances des extensions de E_1 et $\overline{E_1}$ dans P_1 et de F_2 et $\overline{F_2}$ dans P_2 (cf. Section 1.1), nous calculons l'ensemble des observations noté $Ob(E_1, \overline{E_1}, F_2, \overline{F_2})$ et défini de la façon suivante :

Définition 2.1

$$Ob(E_1, \overline{E_1}, F_2, \overline{F_2}) = (Card(E_1, \overline{E_1}, F_2, \overline{F_2}), MD(E_1, \overline{E_1}, F_2, \overline{F_2}))$$

où $Card(E_1, \overline{E_1}, F_2, \overline{F_2})$ est le quadruplet des nombres d'instances de E_1 et $\overline{E_1}$ dans P_1, et de F_2 et $\overline{F_2}$ dans P_2 : $(|ext(E_1, P_1)|, |ext(\overline{E_1}, P_1)|, |ext(F_2, P_2)|, |ext(\overline{F_2}, P_2)|)$, et $MD(E_1, \overline{E_1}, F_2, \overline{F_2})$ est une matrice à 4 colonnes et m lignes où m est le nombre d'attributs booléens décrivant les méta-données (cf. Section 1.1). La ligne k notée

$MD[k]$ est le quadruplet $(e_1^k, \overline{e_1}^k, f_2^k, \overline{f_2}^k)$ des nombres d'instances de E_1, $\overline{E_1}$, F_2 et $\overline{F_2}$ (dans P_1 et P_2 respectivement) ayant l'attribut numéro k valant 1 dans leur vecteur de méta-données booléennes.

2.1 Modélisation de la probabilité d'un mapping

Nous définissons la probabilité d'un mapping $E_1 \sqsubseteq F_2$ comme la probabilité d'appartenir à $\overline{E_1} \cup F_2$ sachant les observations $Ob(E_1, \overline{E_1}, F_2, \overline{F_2})$

Définition 2.2
Soit X_{E_1} (respectivement X_{F_2}) la variable aléatoire binaire définie sur l'ensemble union des ressources de P_1 et P_2 par $X_{E_1}(r)$ = 1 ssi $r \in ext(E_1, P_1)$ (respectivement $X_{F_2}(r)$ = 1 ssi $r \in ext(B_2, P_2)$). La probabilité du mapping $E_1 \sqsubseteq F_2$ est notée $P(E_1 \sqsubseteq F_2)$ et est définie par :

$$P(E_1 \sqsubseteq F_2) = P(X_{E_1} = 0 \; ou \; X_{F_2} = 1 | Ob(E_1, \overline{E_1}, F_2, \overline{F_2}))$$

On peut définir de la même façon la probabilité d'une inclusion entre classes d'une même ontologie.
Le théorème suivant établit le lien entre la sémantique logique et la sémantique probabiliste des mappings. Il montre que la fonction de probabilité entre mappings est monotone par rapport à l'implication logique entre mappings. Il repose sur l'hypothèse de cohérence des méta-données observées par rapport aux ontologies de chaque pair.

Définition 2.3
Soit O_i l'ontologie d'un pair P_i, les méta-données observées sont cohérentes par rapport à O_i si pour tout axiome d'inclusion $E_i \sqsubseteq E_i'$ de O_i :
$P(E_i \sqsubseteq E_i' | Ob(E_i, \overline{E_i}, E_i', \overline{E_i'})) = 1$

Théorème 2.1
Soient 2 mappings $E_1 \sqsubseteq F_2$ et $E_1' \sqsubseteq F_2'$ entre 2 ontologies O_1 et O_2 de 2 pairs P_1 et P_2. A condition que les méta-données observées dans chaque pair P_1 et P_2 soient cohérentes par rapport aux ontologies respectives O_1 et O_2, on a :

$$Si \; O_1 \cup O_2, E_1 \sqsubseteq F_2 \models E_1' \sqsubseteq F_2' \; alors \; P(E_1 \sqsubseteq F_2) \leq P(E_1' \sqsubseteq F_2')$$

2.2 Estimation bayésienne de la probabilité d'un mapping

Nous présentons d'abord un théorème donnant une formule exprimant la probabilité d'un mapping par rapport aux observations, puis nous donnons la façon dont on peut estimer chacun de ses membres par une méthode statistique bayésienne. Nous abrégeons désormais $Ob(E_1, \overline{E_1}, F_2, \overline{F_2})$ par $Ob_{1,2}$, l'égalité $X_{E_1} = 1$ par E_1, et $X_{F_2} = 0$ par $\overline{F_2}$.
Soit X_i pour $1 \leq i \leq m$ la variable aléatoire binaire définie sur l'ensemble union des ressources de P_1 et P_2, par $X_i(r) = 1$ ssi l'attribut booléen i vaut 1 dans les méta-données booléennes de r. On fait les hypothèses suivantes :

– **(h1)** X_{E_1} et X_{F_2} sont indépendantes conditionnellement à $O_{1,2}$ et aux X_i : i.e. dès qu'on connaît $Ob_{1,2}$ et les X_i, le fait de savoir qu'une instance appartient à E_1 ou non n'apporte pas plus d'information sur la probabilité qu'elle appartienne à F_2
– **(h2)** les X_i sont indépendantes conditionnellement aux classes (hypothèse classique en apprentissage naïve bayes) et indépendantes

Théorème 2.2
*Sous les hypothèses **(h1)** et **(h2)** :* $P(E_1 \sqsubseteq F_2) =$

$$1 - P(E_1|Ob_{1,2})P(\overline{F_2}|Ob_{1,2}) \prod_{i=1}^{m} \sum_{v_i=0}^{1} \frac{P(X_i = v_i|E_1, Ob_{1,2})P(X_i = v_i|(\overline{F_2}, Ob_{1,2})}{P(X_i = v_i|Ob_{1,2})} \quad \textbf{(1)}$$

Malgré la complexité apparente de cette formule, son calcul ne nécessite qu'un nombre d'opération en $O(m)$ à partir de ses opérandes. Le théorème 2.3 fournit un moyen d'estimer la valeur des paramètres de ces opérations. Ci-dessous figurent ces $2 + 3m$ paramètres à estimer :
– $P(E_1|Ob_{1,2})$ et $P(\overline{F_2}|Ob_{1,2})$: probabilités des classes E_1 et $\overline{F_2}$ sachant les observations
– $P(X_i = 1|Ob_{1,2})$ pour $1 \leq i \leq m$: probabilité qu'une instance ait l'attribut i à 1 dans ses méta-données booléennes sachant les observations
– $P(X_i = 1|E_1, Ob_{1,2}), P(X_i|\overline{F_2}, Ob_{1,2})$ pour $1 \leq i \leq m$: probabilité qu'une instance ait la valeur 1 pour l'attribut i dans ses méta-données booléennes *sachant qu'elle appartient à E_1, ou à F_2, connaissant les observations sur P_1 et P_2.*

Nous montrons maintenant comment estimer ces paramètres nécessaires au calcul de $P(E_1 \sqsubseteq F_2)$, par une approche bayésienne de la statistique (Schervish, 2002). Afin d'alléger les formules dans cette partie, on notera :
– $n_{E_1} = |ext(E_1, P_1)|$ et $n_{F_2} = |ext(F_2, P_2)|$
– n_{P1} nombre d'instances du pair P_1, n_{P2} celui du pair P_2.
– $n = n_{P1} + n_{P2}$ le nombre d'instances sur les pairs P_1 et P_2
– $n_i = a_i + \overline{a_i} + b_i + \overline{b_i}$ le nombre d'instances u de P_1 et P_2 avec $mdb(u)[i] = 1$
On s'intéresse en premier lieu à l'estimation de $P(E_1|Ob_{1,2})$. Le résultat de l'expérience consistant à prendre au hasard une instance et à regarder si elle appartient à E_1 suit une loi de Bernouilli de paramètre $p = P(A_1|Ob_{1,2})$. C'est la loi probabiliste la plus simple, correspondant à 2 états, de probabilités respectives p et $1 - p$. L'approche bayésienne consiste à modéliser le paramètre recherché p comme une variable aléatoire.

Théorème 2.3
Si $X_{E_1}|Ob_{1,2}$ suit une loi de Bernouilli de paramètre $p = P(X_{E_1}|Ob_{1,2})$, on peut estimer p de façon bayésienne par la formule suivante (Schervish, 2002; Gia-Hien Nguyen, 2008) :

$$P(E_1|Ob_{1,2}) = p \approx \frac{1 + n_{E_1}}{2 + n_{E_1} + (n_{P1} - n_{E_1})} = \frac{1 + n_{E_1}}{2 + n_{P1}}$$

En transposant ce théorème aux autres paramètres, on obtient les formules d'estimation suivantes :
– $P(\overline{F_2}|Ob_{1,2}) \approx \frac{1+n_{P2}-n_{F_2}}{2+n_{P2}}$

– $P(X_i = 1|Ob_{1,2}) \approx \frac{1+n_i}{2+n}$, et $P(X_i = 0|Ob_{1,2}) \approx= 1 - \frac{1+n_i}{2+n}$
– $P(X_i = 1|E_1, Ob_{1,2}) \approx \frac{1+n_i+a_i}{2+n+n_{E_1}}$, et idem que ci-dessus pour $X_i = 0$
– $P(X_i = 1|\overline{F_2}, Ob) \approx \frac{1+n_i+\overline{b_i}}{2+n+n_{P2}-n_{F_2}}$ (idem pour $X_i = 0$)

Le théorème pour l'estimation de $P(E_1 \sqsubseteq F_2)$ est un corollaire des théorèmes 2.2 et 2.3 :

Théorème 2.4
Sous les hypothèses du théorème 2.2, en utilisant l'estimation bayésienne du théorème 2.3, on estime $P(E_1 \sqsubseteq F_2)$ sachant les observations $Ob_{1,2}$ issues des méta-données par :

$$P(E_1 \sqsubseteq F_2) \approx 1 - \frac{1+n_{E_1}}{2+n_{P1}} \frac{1+n_{P2}-n_{F_2}}{2+n_{P2}} \prod_{i=1}^{m} \left(\frac{t_E^i t_F^i}{t_X^i} + \frac{(1-t_E^i)(1-t_F^i)}{1-t_X^i} \right)$$

avec $t_E^i = \frac{1+n_i+a_i}{2+n+n_{E_1}}$, $t_F^i = \frac{1+n_i+\overline{b_i}}{2+n+n_{P2}-n_{F_2}}$, $t_X^i = \frac{1+n_i}{2+n}$

Cette estimation sera notée $P_{estim}(E_1 \sqsubseteq F_2, Ob_{1,2})$.

Malgré l'apparence complexe de cette formule, chacun de ses membres se calculent facilement à partir des n_{P1}, n_{P2}, n_{E_1} et n_{F_2}, a_i et b_i, $\overline{a_i}$ et $\overline{b_i}$ contenus dans $Ob_{1,2}$, eux-mêmes obtenus en comptant les instances de P_1 et P_2 respectant les critères adéquats. On remarque que si E_1 n'a pas d'instance (dans P_1), le calcul se réduit à l'estimation suivante : $P(E_1 \sqsubseteq F_2) \approx \frac{1+n_{E_1}}{2+n_{P1}} \frac{1+n_{P2}-n_{F_2}}{2+n_{P2}}$ qui est celle de $1-P(E_1|Ob_{1,2})P(\overline{F_2}|Ob_{1,2})$, le produit faisant 1, indépendants des statistiques sur les valeurs des attributs. L'absence d'instances dans E_1 se traduit donc de façon cohérente par la non-prise en compte de la corrélation entre valeurs des attributs dans les classes concernées par le mapping.

3 Enumération et test d'un ensemble de mappings candidats

Le problème abordé dans cette section consiste à énumérer et tester (i.e. savoir si $P(m) > s$, en réalité si $P_{estim}(m, Ob) > s$) de façon efficace un grand ensemble de mappings candidats $\mathcal{M}$ entre deux ontologies O_1 et O_2. Pour éviter de tester successivement tous les mappings de $\mathcal{M}$ par la méthode décrite à la section 2, on structure $\mathcal{M}$ par une relation d'ordre fondée sur la sémantique, puis on exploite la propriété de monotonie de la probabilité par rapport à cet ordre, qui découle du théorème 2.1.

3.1 Ordre sur $\mathcal{M}$ et monotonie de la probabilité des mappings

Nous introduisons tout d'abord la relation d'équivalence logique $\equiv_{O_1,O_2}$ suivante :
m et m' sont équivalents ssi $O_1, O_2 \models m \Leftrightarrow m'$.
La relation $\preceq$ est la relation d'implication logique entre mappings compte tenu des ontologies :

$$m \preceq m' \quad \Leftrightarrow \quad O_1, O_2, m \models m'$$

$m \preceq m'$ est une relation d'ordre sur l'ensemble $\mathcal{M}' = \mathcal{M}/\equiv_{O_1,O_2}$, à savoir l'ensemble des mappings candidats quotienté par l'équivalence.
Le théorème 2.1 implique directement la propriété de monotonie suivante :

Théorème 3.1
Si les méta-données observées dans chaque pair P_1 et P_2 sont cohérentes par rapport aux ontologies respectives O_1 et O_2 alors : $m \preceq m' \ \Rightarrow \ P(m) \leq P(m')$.

Ce théorème entraîne deux conséquences (équivalentes) sur lesquelles est basé l'algorithme 1 : Sous l'hypothèse de la cohérence des méta-données par rapport aux ontologies O_1 et O_2, étant donnés deux mappings m et m' entre O_1 et O_2, avec $m \preceq m'$, et un seuil quelconque $0 \leq s \leq 1 : P(m) > s \Rightarrow P(m') > s \ \ et \ \ P(m') < s \Rightarrow P(m) < s$

Influence de la modélisation de la probabilité d'un mapping sur la monotonie
On peut penser à modéliser $P(E_1 \sqsubseteq F_2)$ par une probabilité conditionnelle :
$P(E_1 \sqsubseteq F_2) = P(X_{F_2} = 1 | X_{E_1} = 1, Ob)$
Dans ce cas, il n'y a pas de monotonie. De plus, cette modélisation manque de cohérence avec la logique car l'équivalence de la contraposition n'est pas préservée au niveau probabiliste : $P(E_1 \sqsubseteq F_2) \neq P(\overline{F_2} \sqsubseteq \overline{E_1})$

3.2 Algorithme d'énumération et de test

L'entrée de l'algorithme 1 est constituée :
– des observations Obs : $Obs(E_1 \sqsubseteq F_2) = Ob(E_1, \overline{E_1}, F_2, \overline{F_2})$ pour chaque mapping $E_1 \sqsubseteq F_2$
– G le graphe de la réduction transitive de la relation $\preceq$ sur $\mathcal{M}'$. On trouve un algorithme à cet effet dans (Schwarz, 2007).
– un seuil s
La sortie de cet algorithme est l'ensemble des minimaux de l'ensemble des mappings de $\mathcal{M}$ dont la probabilité (estimée) dépasse s.
On suppose qu'on a déjà vérifié la cohérence des méta-données par rapport aux ontologies, à partir de l'estimation des probabilités des liens de subsomption avec la méthode de la section 2. L'algorithme utilise les primitives suivantes :
– INF(m, G) et SUP(m, G) : retournent respectivement les ensembles de mappings inférieurs et supérieurs de m par rapport à $\preceq$ dans $\mathcal{M}'$
– MAX(G) : maximaux des mappings de $\mathcal{M}'$
– PRED(m, G) et SUCC(m, G) : prédécesseurs et successeurs de m dans les mappings candidats, par rapport à $\preceq$. m' est prédécesseur de m ssi $m' \preceq m$ et s'il n'existe pas $m'' \in \mathcal{M}'$ tel que $m' \preceq m'' \preceq m$. Successeur est la relation inverse de prédécesseur.
Le principe général de l'algorithme est le parcours d'un ensemble de mappings courants $Current \subseteq \mathcal{M}'$, et la construction de la liste suivante $Next$ en fonction des résultats du parcours de $Current$. S'ils dépassent le seuil, les mappings sont stockés dans M_{Val}, sinon dans M_{NVal}.
$Current$ est initialisé avec les maximaux de $\mathcal{M}'$. Chaque mapping de $Current$ non encore testé est soumis au test $P_{estim}(m, Obs(m)) > s$ (ligne 7) :

– Si $P_{estim}(m, Ob) > s$, alors on l'ajoute ainsi que ses mappings supérieurs de $\mathcal{M}'$ dans M_{Val} (lignes 11 et 12). On ajoute aussi tous les prédécesseurs de m dans $Next$ (ligne 13), car on ne sait rien sur eux (sauf si on en a déjà vu avant, auquel cas on épure $Next$ ligne 18).
– Sinon, on ajoute m et tous les mappings inférieurs à m dans M_{NVal} (l. 15).

Algorithm 1 Enumération et test d'un ensemble de mappings candidats

Require: Obs, G, s
Ensure: Retourne les minimaux de l'ensemble des mappings de $\mathcal{M}/\equiv_{O_1,O_2}$ dont la probabilité dépasse s

1: $M_{Val} \leftarrow \emptyset$, $M_{NVal} \leftarrow \emptyset$
2: $M_{NotMin} \leftarrow \emptyset$
3: $Current \leftarrow \text{MAX}(G)$
4: **while** $Current \neq \emptyset$ **do**
5: $Next \leftarrow \emptyset$
6: **for** each $m \in Current$ **do**
7: **if** $m \notin M_{Val}$ **and then** $P_{estim}(m, Obs(m)) > s$ **then**
8: $E \leftarrow \text{SUP}(m, G)$
9: $M_{Val} \leftarrow M_{Val} \cup E$
10: $M_{NotMin} \leftarrow M_{NotMin} \cup E \setminus \{m\}$
11: **end if**
12: **if** $m \in M_{Val}$ **then**
13: $Next \leftarrow Next \cup \text{PRED}(m, G)$
14: **else**
15: $M_{NVal} \leftarrow M_{NVal} \cup \text{INF}(m, G)$
16: **end if**
17: **end for**
18: $Current \leftarrow Next \setminus M_{NVal}$
19: **end while**
20: Return($M_{Val} \setminus M_{NotMin}$)

La monotonie est doublement exploitée : elle sert d'une part à construire le niveau suivant, et d'autre part à propager vers les supérieurs ou inférieurs d'un mapping le fait que sa probabilité dépasse le seuil s ou non. Nous avons montré la correction de cet algorithme qu'exprime le théorème suivant :

Théorème 3.2
L'algorithme 1 termine, et soit $Result = M_{Val} \setminus M_{NotMin}$ l'ensemble qu'il retourne $Result$ est l'ensemble des mappings minimaux par rapport à $\preceq$ de l'ensemble des mappings de $\mathcal{M}'$ dont la probabilité dépasse le seuil s.

4 Conclusion

Par rapport aux travaux existants sur l'alignement d'ontologies, nous nous positionnons au sein des méthodes d'alignement d'ontologies à bases d'instances (e.g. Doan

et al. (2002), fondé sur une méthode de classification). Nous ne faisons pas l'hypothèse de connaissance centralisée. L'originalité de notre approche est la prise en compte de méta-données normalisées, transformées (sans l'avoir détaillé ici) en logique propositionnelle, qui permettent d'estimer de façon bayésienne la sémantique probabiliste de mappings. Celle-ci est différente de celle de (Dong *et al.*, 2007) qui définit des sémantiques pour des mappings probabilistes (p-mappings) dans le cadre d'intégration de schémas relationnels à base de correspondances entre attributs, ainsi que la consistance des données par rapport aux p-mappings. Les p-mappings et les probabilités sont supposés fournis par une méthode en amont.
En outre, nous avons fourni une méthode et un algorithme pour structurer un ensemble de mappings candidats afin de fournir en sortie les mappings dont la probabilité d'un mapping dépasse un certain seuil, de manière efficace. Pour celà, nous exploitons la monotonie de la probabilité sur un ordre fondé sur la sémantique logique.
Nous avons comme perspective la mise en application sur des jeux de tests générés et réels, afin de mesurer les performances en terme de complexité et de robustesse, en variant les méthodes d'estimation. Nous étudierons le prétraitement des méta-données. Le langage de mapping pourra être étendu, notamment avec des expressions disjonctives et conjonctives.

Références

ADJIMAN P., CHATALIC P., GOASDOU F., ROUSSET M.-C. & SIMON L. (2005). SomeWhere in the Semantic Web. In *International Workshop on Principles and Practice of Semantic Web Reasoning*.

DOAN A., MADHAVAN J., DOMINGOS P. & HALEVY A. (2002). Learning to map between ontologies on the semantic web. In *WWW '02 : Proceedings of the 11th international conference on World Wide Web*, p. 662–673, New York, NY, USA : ACM.

DONG X. L., HALEVY A. Y. & YU C. (2007). Data integration with uncertainty. In *VLDB*, p. 687–698.

GIA-HIEN NGUYEN, PHILIPPE CHATALIC M.-C. R. (2008). A Probabilistic Trust Model for Semantic Peer to Peer systems.

RUSSELL S., NORVIG P., MICLET L. & POPINEAU F. (2006). *Intelligence Artificielle. 2e édition*. PEARSON EDUCATION.

SCHERVISH D. G. (2002). *Probability and Statistics, p. 336*. Addison Wesley.

SCHWARZ U. M. (2007). Transitive reduction and Union Find.

SHVAIKO P. & EUZENAT J. (2005). A survey of schema-based matching approaches. p. 146–171.

TOM M. MITCHELL M.-H. (1997). *Machine Learning, by Tom M. Mitchell, McGraw-Hill*.