Une aide à la découverte de mappings dans SomeRDFS

François-Élie Calvier, Chantal Reynaud

LRI, Univ Paris-Sud & INRIA Futurs
4, rue Jacques Monod - Bât. G
91893 Orsay Cedex
francois.calvier@lri.fr,
chantal.reynaud@lri.fr
http://www.lri.fr/iasi

Résumé. Dans cet article, nous nous intéressons à la découverte de mises en correspondance entre ontologies distribuées modélisant les connaissances de pairs du système de gestion de données P2P SomeRDFS. Plus précisément, nous montrons comment exploiter les mécanismes de raisonnement mis en œuvre dans SomeRDFS pour aider à découvrir des mappings entre ontologies. Ce travail est réalisé dans le cadre du projet MediaD en partenariat avec France Telecom R&D.

1 Introduction

Nous nous intéressons à la découverte de correspondances, ou mappings, entre ontologies distribuées modélisant les connaissances de pairs du système de gestion de données P2P (PDMS) SomeRDFS. Un PDMS est un système constitué de pairs autonomes qui communiquent pour répondre collectivement à une requête. Les communications entre pairs s'établissent grâce à des mappings qui définissent des relations sémantiques entre leurs connaissances. Un PDMS est sollicité via l'interrogation d'un des pairs qui pourra ensuite faire appel aux autres pour répondre. Une spécificité des PDMS est que chaque pair ne connaît que ses propres connaissances et les mappings le connectant à d'autres pairs. Dans ce cadre, nous cherchons à augmenter le nombre de mappings de chaque pair afin d'améliorer les réponses fournies globalement par le système, en quantité et en qualité.

Nous travaillons, dans le cadre du projet MediaD (projet financé par France Telecom R&D), dont l'objectif est la création d'un environnement déclaratif de construction de systèmes de gestion de données P2P. Ces travaux ont conduit au développement de la plate-forme SomeRDFS (Adjiman et al., 2006) au sein de laquelle nous situons notre travail.

Nous présenterons dans un premier temps le contexte de notre travail. Nous montrerons ensuite comment les requêtes des utilisateurs peuvent être exploitées pour identifier des raccourcis de mappings ainsi que des relations cibles à partir desquelles des mises en correspondances intéressantes peuvent être trouvées. Étant données ces relations cibles, nous proposerons alors des techniques basées sur l'interrogation du système pour construire des ensembles de candidats à un mapping. Nous présenterons ensuite quelques travaux proches. Enfin, nous conclurons et présenterons quelques perspectives.

2 Contexte de travail

Dans le contexte de SomeRDFS, les ontologies et les mappings sont exprimés en RDF(S) tandis que les données sont représentées en RDF. Il est ainsi possible de définir des classes, des sous-classes, des propriétés, des sous-propriétés, de typer le domaine et le co-domaine des propriétés. Les constructeurs autorisés sont l'inclusion de classes, l'inclusion de propriétés, le typage de domaine et de co-domaine. Ce langage est basé sur des relations unaires qui représentent les classes et des relations binaires qui représentent des propriétés.

Les ontologies des pairs sont représentées à l'aide d'expressions RDFS composées uniquement de relations appartenant au vocabulaire du pair. Nous notons $\mathcal{P}$:R la relation R (classe ou propriété) de l'ontologie du pair $\mathcal{P}$.

Les données d'un pair sont associées à des relations faisant partie de son vocabulaire. Un mapping correspond à une inclusion entre classes ou propriétés de 2 pairs différents ou au typage d'une propriété d'un pair donné avec une classe d'un autre pair (cf. TAB. 1). Ainsi les mappings sont également des expressions RDF(S). Leur spécificité est d'être construites à partir du vocabulaire des ontologies de pairs différents qui établissent ainsi des correspondances sémantiques entre eux. Chaque pair $\mathcal{P}$ peut être sollicité à l'aide de requêtes exprimées avec son propre vocabulaire.

Mappings	Notation LD	Traduction en logique du premier ordre
Inclusion de classes	$\mathcal{P}_1{:}C_1 \sqsubseteq \mathcal{P}_2{:}C_2$	$\forall X, \mathcal{P}_1{:}C_1(X) \Rightarrow \mathcal{P}_2{:}C_2(X)$
Inclusion de propriétés	$\mathcal{P}_1{:}P_1 \sqsubseteq \mathcal{P}_2{:}P_2$	$\forall X \forall Y, \mathcal{P}_1(X,Y) \Rightarrow \mathcal{P}_2(X,Y)$
Typage de domaine d'une propriété	$\exists \mathcal{P}_1{:}P \sqsubseteq \mathcal{P}_2{:}C$	$\forall X \forall Y, \mathcal{P}_1{:}(X,Y) \Rightarrow \mathcal{P}_2{:}C(X)$
Typage de co-domaine d'une propriété	$\exists \mathcal{P}_1{:}P^- \sqsubseteq \mathcal{P}_2{:}C$	$\forall X \forall Y, \mathcal{P}_1{:}(X,Y) \Rightarrow \mathcal{P}_2{:}C(Y)$

TAB. 1 – *Mappings*

Le calcul des réponses aux requêtes se fait en deux temps. Les requêtes sont d'abord réécrites en un ensemble de requêtes les subsumant. Le calcul des réécritures maximales de chaque atome d'une requête fournit un ensemble de conjonctions de relations (classes ou propriétés). Les requêtes sont ensuite propagées du fait de l'existence de mappings avec des relations d'autres pairs. Ces derniers transmettront les réécritures obtenues. Ces réécritures seront ensuite évaluées afin d'obtenir les données associées.

3 Exploitation des réponses aux requêtes utilisateurs

3.1 Découverte de raccourcis de mappings

Un raccourci de mappings est un mapping résultant de la composition de mappings existants. Les raccourcis de mappings renforcent le réseau en créant des liens directs entre des relations de deux pairs différents là où jusqu'alors n'existaient, dans le PDMS, que des liens indirects. L'objectif n'est pas de rajouter ce type de mapping systématiquement mais de façon sélective. En effet ces mappings peuvent être intéressants à représenter car, bien qu'ils ne permettent pas d'aboutir à des réponses plus riches, ils peuvent être très utiles en cas de disparition de pairs du PDMS. La découverte des raccourcis de mappings à repésenter peut être automatisée. Nous proposons, pour cela, d'exploiter le mécanisme de réponse aux requêtes

des utilisateurs et d'appliquer ensuite des techniques de sélection des raccourcis pertinents à représenter.

Concernant le traitement des requêtes utilisateurs aboutissant à l'obtention des données, nous proposons de dissocier les deux étapes du traitement qui sont : (1) le calcul des réécritures maximales de chaque atome de la requête posée, (2) l'évaluation des réécritures. Cette dissociation est intéressante lorsqu'un utilisateur ne trouve pas, au sein du vocabulaire du pair auquel il s'adresse, la relation lui permettant de définir précisément les données qu'il recherche, et qu'il est dans l'obligation d'indiquer une autre relation. Cette autre relation peut être plus spécifique. Dans ce cas, toutes les réponses attendues ne seront pas obtenues. Elle peut être plus générale. Dans cet autre cas, elle permettra d'obtenir toutes les réponses attendues par l'utilisateur mais contiendra, en revanche, également des réponses inutiles. Ainsi, par exemple, si l'utilisateur s'adresse au pair $\mathcal{P}_1$ dans l'espoir d'obtenir les données de la classe $SteelSculptor$, il peut, en l'absence de la classe $SteelSculptor$ dans $\mathcal{P}_1$, s'intéresser aux données d'une classe plus générale en posant la requête $Q_2(X) \equiv \mathcal{P}_1{:}Artist(X)$. Il obtiendra, parmi les réécritures, $\mathcal{P}_2{:}SteelSculptor(X)$ indiquant que les données instanciant $SteelSculptor(X)$ sont des artistes mais également $\mathcal{P}_2{:}WoodSculptor(X)$ ou $\mathcal{P}_2{:}GlassSculptor(X)$ dont il n'est pas utile de rechercher les données, compte tenu du besoin précis de l'utilisateur. Le fait de dissocier le calcul des réécritures de leur évaluation est, dans ce cas, intéressant. Il permet à l'utilisateur de sélectionner les réécritures pour lesquelles il demande l'évaluation.

Nous proposons de stocker les raccourcis de mappings correspondant à des liens directs avec des relations qualifiées d'intéressantes pour les utilisateurs. Ces relations, dites intéressantes, appartiennent au vocabulaire de pairs distants. Elles sont plus spécifiques que celles du pair étudié. Elles n'apparaissent pas, pour l'instant, dans ses mappings, mais sont apparues dans des réécritures et les utilisateurs ont, à plusieurs reprises, demandé leur évaluation. L'ensemble des mappings potentiels ainsi stockés seront traités globalement ultérieurement pour sélectionner ceux qui seront effectivement ajoutés.

3.2 Identification de relations cibles

L'intérêt principal des mappings est de permettre de propager des requêtes à des pairs distants pour qu'ils contribuent aux réponses. Lorsqu'un utilisateur s'adresse à un pair, il arrive toutefois que les réponses proviennent toutes de ce pair. Cette situation révèle un manque de mappings de spécialisation entre les relations de ce pair et celles de pairs distants. L'identification de ce phénomène est possible par analyse des réponses obtenues aux requêtes utilisateurs, ou, plus précisément, par observation de la localisation des éléments qui composent les réponses.

Une étude nous a permis de définir deux cas pour lesquels le calcul de réécritures est un indice pour trouver des relations appartenant à des pairs distants qui peuvent être rapprochées de relations du pair interrogé. Nous présentons successivement ces deux cas en nous limitant aux classes, le fonctionnement étant similaire pour les propriétés.

Cas 1 (cf. FIG 1) :

Soient les pairs $\mathcal{P}_1$, $\mathcal{P}_2$ et $\mathcal{P}_3$ contenant respectivement les classes C_1, C_2 et C_3 et les mappings suivants : $\mathcal{P}_1{:}C1(X) \Rightarrow \mathcal{P}_2{:}C_2(X)$ et $\mathcal{P}_3{:}C_3(X) \Rightarrow \mathcal{P}_2{:}C_2(X)$ représentés chacun dans les deux pairs concernés.

1. L'utilisateur s'adresse à $\mathcal{P}_3$ et pose la requête $Q_4(X) \equiv \mathcal{P}_3{:}C_3(X)$
2. SomeRDFS ne fournit aucune réécriture.

Cas 2 (cf. FIG 1) :
Soient le pair $\mathcal{P}_1$ contenant la classe $C1$ et le pair $\mathcal{P}_2$ contenant les classes C_2 et C_3 telles que $\mathcal{P}_2{:}C_2(X) \Rightarrow \mathcal{P}_2{:}C_3(X)$. Nous supposerons que le mapping $\mathcal{P}_2{:}C2(X) \Rightarrow \mathcal{P}_1{:}C1(X)$ est représenté à la fois dans $\mathcal{P}_1$ et dans $\mathcal{P}_2$.

1. L'utilisateur s'adresse à $\mathcal{P}_2$ et pose la requête $Q_5(X) \equiv \mathcal{P}_2{:}C_3(X)$.

2. SomeRDFS fournit la réécriture suivante : $\mathcal{P}_2{:}C_2(X)$. Aucune réécriture composée de relations de pairs distants n'est donnée.

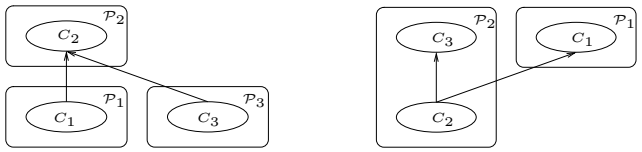

FIG. 1 – *Cas 1 et Cas 2*

C_3 dans le cas 1 et C_2 dans le cas 2 sont respectivement appelées relations cibles car c'est à partir d'elles que des mises en correspondance intéressantes peuvent être trouvées comme nous le montrons dans la section qui suit. Les deux cas présentés ci-dessus sont des cas élémentaires qui, s'ils sont combinés, peuvent permettre de traiter des cas plus complexes. La notion de relation cible pourra être étendue à toute relation locale apparaissant dans une réécriture et généralisant une relation cible.

4 Identification de candidats pour la mise en correspondance

Nous proposons une procédure d'identification de candidats à un mapping, qui s'appuie sur les relations cibles, notées R_{CIBLE}, supposées appartenir au pair $\mathcal{P}_{CIBLE}$. Sans liens de spécialisation avec d'autres pairs via des mappings, ces relations ne permettent pas de propager le raisonnement vers un autre pair. Pour chaque relation cible, nous nous proposons alors de déterminer, dans un second temps, un ensemble des relations entre lesquelles il serait pertinent de rechercher des mises en correspondance. Cet ensemble de relations sera appelé candidats au mapping et noté CM. Le processus de découverte de mappings nouveaux est donc un processus en trois étapes : (1) recherche des relations cibles, (2) recherche d'ensembles de candidats au mapping, (3) alignement des éléments de l'ensemble de candidats au mapping.

Notre approche pour la recherche d'ensembles de candidats au mapping est basée sur l'idée selon laquelle il est pertinent de rechercher des mises en correspondance entre des relations ayant des points communs. Dans le cadre de SomeRDFS, les points communs considérés seront (1) l'existence d'une relation commune plus générale, ou bien (2) l'existence d'une relation commune plus spécifique. Nous proposons un processus de recherche de candidats au mapping pour chacune de ces situations. Nous nous basons sur l'existence d'une relation commune plus générale lorsque R_{CIBLE} n'a qu'une relation R_g (au sein du même pair ou dans la définition d'un mapping) qui la généralise. Dans ce cas, nous proposons de poser une requête sur cette relation ($Q(X) \equiv R_g(X)$) et d'en calculer les réécritures. Ces

dernières fournissent un ensemble de relations plus spécifiques que R_g entre lesquelles il est pertinent de rechercher l'existence de correspondances. Cet ensemble constitue CM. Si $\exists! R_g/R_{CIBLE} \Rightarrow R_g$ et $Q(X) \equiv R_g(X)$ alors CM = l'ensemble des relations correspondant à des réécritures de $Q(X)$. Nous nous basons sur l'existence d'une relation commune plus spécifique lorsque R_{CIBLE} a plusieurs relations R_g (au sein du même pair ou dans la définition d'un mapping) qui la généralisent. Dans ce cas, l'ensemble de ces généralisants constitue un ensemble de candidats au mapping CM. Si $G = \{R_g/R_{CIBLE} \Rightarrow R_g\}$ avec $G = G_1 \bigcup G_2/G_1 = \{R_g \mid R_g \in \mathcal{P}_{CIBLE}\}$ et $G_2 = \{R_g \mid R_g \notin \mathcal{P}_{CIBLE}\}$ alors $CM = G$. Ces deux situations correspondent aux cas 1 et 2 décrits dans la section précédente.

La recherche de mappings devra ensuite être effectuée entre les éléments de l'ensemble CM. Cet ensemble comprend des relations de différents pairs, celui à qui la requête a été posée pour obtenir cet ensemble, et les pairs distants ayant contribué à la réponse à la requête via des réécritures. Le pair interrogé a une certaine compréhension des relations de CM faisant partie de son vocabulaire puisque celles-ci font partie de son ontologie. En revanche, les relations de CM appartenant au vocabulaire de pairs distants sont inconnues du pair interrogé. Pour lui, il s'agit de noms de relations isolées. Le problème d'alignement qui se pose consiste alors à mettre en correspondance des relations prises isolément avec des relations appartenant à une ontologie. Le travail que nous avons réalisé jusqu'alors permet d'isoler les ensembles de relations à aligner. Il doit être complété par l'application de techniques d'alignement appropriées. L'étude des techniques les plus adaptées fait partie de nos perspectives.

5 Travaux proches

Appliqué aux systèmes P2P, le problème d'alignement peut être résolu de différentes façons. Certains travaux comme Piazza (Halevy et al., 2004) proposent l'application de techniques éprouvées testées dans le cadre de systèmes d'intégration mais supposent, pour l'alignement, que les ontologies de tous les pairs sont connues de tous, ou du moins accessibles dans leur totalité par tous. D'autres travaux font intervenir des ontologies dites de référence avec lesquelles l'ontologie de chaque pair peut être liée, ce qui évite la mise en relation directe des ontologies des pairs les unes avec les autres (Herschel et Heese, 2005). Enfin, une troisième catégorie de travaux consiste à étudier le problème de l'alignement d'ontologies lorsque toutes les ontologies composant le système P2P sont distribuées et qu'il en existe aucune qui soit connue de tous et qui puisse servir d'ontologie de référence. Ce problème a été étudié dans le système Helios (Castano et al., 2003) où la découverte de mappings est basée sur l'interrogation des pairs du réseau.

Par rapport à cet état de l'art, l'approche que nous proposons dans ce papier est spécifique au contexte distribué, tous les pairs étant considérés de la même façon et les ontologies étant réparties entre tous les pairs du système. En ce sens, elle se rapproche de la 3ème catégorie de travaux décrite ci-dessus. Nous exploitons toute la richesse du raisonnement pouvant être mis en oeuvre au sein d'un PDMS, suite à l'envoi de requêtes. Toutefois, contrairement aux travaux de (Castano et al., 2003), les requêtes ne sont propagées qu'à un nombre réduit de pairs, elles ne sont pas envoyées à l'ensemble des pairs du PDMS. Par ailleurs, les requêtes exploitées pour la découverte de mappings ont toutes la forme de requêtes utilisateurs. Elles ne nécessitent donc pas de concevoir des modules de traitement de requêtes spécifiques. L'originalité de notre approche consiste donc à réutiliser les mécanismes de raisonnement mis en oeuvre

dans SomeRDFS de façon à cibler les éléments à rapprocher puis à réutiliser ou adapter les techniques d'alignement qui ont été expérimentées dans d'autres contextes et dont les résultats se sont avérés être de qualité.

6 Conclusion

Au travers du travail décrit dans ce papier, qui s'inscrit dans le cadre de systèmes P2P, nous avons étudié comment tirer parti du processus de raisonnement logique mis en oeuvre dans SomeRDFS, dans le but d'aider à la découverte de correspondances entre ontologies. Nos perspectives portent sur l'étude des techniques usuelles capables, en particulier, d'aligner des éléments considérés de façon isolée avec d'autres dont on connait précisément l'ontologie à laquelle ils appartiennent (Reynaud et Safar, 2007; Kefi et al., 2006). Nous étudierons également la découverte de mappings prenant appui sur des requêtes ayant la forme de conjonction de relations. Enfin, des tests à plus grande échelle seront réalisés, mettant en jeu un nombre de pairs importants dotés d'ontologies de taille significative.

Références

Adjiman, P., F. Goasdoué, et M.-C. Rousset (2006). SomeRDFS in the semantic web. *Journal on Data Semantics*, p. 158–181.

Castano, S., A. Ferrara, et S. Montanelli (2003). H-match: an algorithm for dynamically matching ontologies in peer-based systems. In *Proc. of the 1st VLDB Int. Workshop on Semantic Web and Databases (SWDB 2003)*, Berlin, Germany.

Halevy, A. Y., Z. G. Ives, D. Suciu, et I. Tatarinov (2004). The piazza peer data management system. *IEEE Transactions on Knowledge and Data Engineering 16*(7), 787–798.

Herschel, S. et R. Heese (2005). Humboldt discoverer: A semantic p2p index for pdms. In *International Workshop Data Integration and the Semantic Web, DISWeb'05, Porto, Portugal, 14/06/05*, http://www.springerlink.com. Springer.

Kefi, H., B. Safar, et C. Reynaud (2006). Alignement de taxonomies pour l'intégration de sources d'information hétérogenes. In *Reconnaissances des Formes et Intelligence Artificielle (RFIA)*.

Reynaud, C. et B. Safar (2007). Techniques structurelles d'alignement pour portail web. In *Numéro spécial fouille du Web, Revue RNTI*.

Summary

This article focuses on mapping distributed ontologies representing knowledge of a peer in SomeRDFS, a peer data management system. We will show how to take advantage of SomeRDFS reasoning in order to help discovering of mappings between ontologies. Our work takes place in the MediaD project.