Passage à l'échelle de la réconciliation de concepts et de la réconciliation de références : quelques points de comparaisons

Nathalie Pernelle*, Fatiha Saïs*

* LRI, Université Paris-Sud 11, F-91405 Orsay Cedex,
INRIA Futurs, 2-4 rue Jacques Monod, F-91893 Orsay Cedex, France
{prenom.nom}@lri.fr

Résumé. L'intégration de données provenant de différentes sources se confronte à deux problèmes majeurs : l'hétérogénéité des schémas et les variations dans les descriptions des instances. Nous avons comparé ces deux problèmes en se focalisant sur l'existence de méthodes permettant de faire face au passage à l'échelle et donc au traitement de données nombreuses et éventuellement distribuées.

1 Introduction

L'intégration de données provenant de différentes sources se confronte à deux problèmes majeurs : l'hétérogénéité des schémas ou des ontologies et les variations dans les descriptions des instances. De nombreux travaux proposent des méthodes permettant de réconcilier des ontologies en découvrant un ensemble de mappings possibles. Il s'agit alors de mettre en relation le vocabulaire de deux ontologies de façon à ce que leur structure mathématique, spécifiée par les axiomes de l'ontologie, soit respectée (Kalfoglou et Schorlemmer (2003)). Selon les approches, les relations possibles peuvent varier : il peut s'agir d'équivalences, de subsumptions, mais aussi de recouvrements ou de disjonctions. D'autres travaux se sont focalisés sur le problème de la réconciliation de références qui découle de l'hétérogénéité dans les descriptions des instances. Dans ce cas, il s'agit de décider si deux descriptions réfèrent ou non à la même entité du monde réel (savoir, par exemple, si deux descriptions décrivent le même hôtel ou la même personne).

Pour décider si deux concepts de deux schémas peuvent être réconciliés, les méthodes peuvent prendre en compte la similarité des concepts qui lui sont liés. De la même façon, pour décider si deux entités peuvent être réconciliés, certaines approches ont une approche globale exploitant les dépendances qui peuvent exister entre réconciliations de références. Par exemple, la réconciliation entre deux références à des articles influe sur la réconciliation entre les deux conférences où sont publiés ces articles.

Spécifier manuellement des mappings entre deux schémas est très coûteux et l'on est confronté au nombre croissant de sources de données en provenance du Web à intégrer. La recherche manuelle de réconciliations de références est encore moins envisageable. Par exemple, les sites comparateurs de prix (e.g. www.kelkoo.com) doivent traiter des millions de références de produits par jour. Ces méthodes doivent donc être aussi automatiques que possible. La mise

à l'échelle des méthodes de réconciliation de concepts et de réconciliations de références se confronte aussi à des contraintes sur le temps d'execution. Certains portails d'information se donnent pour règle de répondre à 30 requêtes en moins de 3 secondes. De plus, les données peuvent être distribuées dans un réseau pair-à-pair.

Le passage à l'échelle de telles méthodes suppose soit la possibilité de disposer d'un ensemble de connaissances permettant de filtrer les données à réconcilier avant de réaliser des traitements plus complexes pour diminuer l'espace des réconciliations possibles, ou de découper le problème en problèmes qui peuvent être résolus de manière indépendante. En ce qui concerne le problème de performance en temps, il est possible de concevoir un système parallèle pour l'exécution des réconciliations ou d'exploiter des méthodes itératives qui permettent de fournir un résultat approximatif après n itérations.

Nous proposons de présenter une comparaison possible des deux problèmes de réconciliation sur ces différents points.

2 Diminuer la taille de l'espace de réconciliation

Pour la réconciliation de références, l'espace de réconciliation représente l'ensemble des couples de références qui sont candidats à la réconciliation. Pour la réconciliation de schéma, l'espace considéré représente l'ensemble des couples de concepts et de relations des schémas (ontologies) à réconcilier. Dans le cas des approches exploitant les instances pour la mise en correspondance de schémas (Euzenat et Valtchev (2004), ou FCA-Merge de Stumme et Maedche (2001)), l'espace de réconciliation des schemas contient également l'ensemble des instances disponibles (appelées aussi objets ou références) des concepts des schémas à réconcilier.

2.1 Filtrage

Pour diminuer la taille de l'espace des réconciliations, les méthodes de réconciliations de références peuvent utiliser en prétraitement des techniques de filtrage. Ces techniques exploitent des connaissances du domaine pour limiter le nombre de couples candidats. Il peut s'agir de :

Méthodes dites de *blocking* : on ne considère que les paires de références qui possèdent une (ou plusieurs) caractéristiques communes, caractéristiques telle que le numéro de ISBN pour les livres, ou encore le nom de famille pour les personnes. Ces techniques ont été introduites par (Newcombe et Kennedy (1962)) et sont utilisées dans des travaux récents tels que Baxter R. (2003). Si il existe une telle caractéristique qui comporte m valeurs, l'espace des réconciliations est divisé par m.

Le corpus CORA (corpus de citations de publications scientifiques qui a servi de benchmark) comporte 6000 références d'articles, de conférences, de journaux et d'auteurs. L'espace de éconciliation concernant les articles et les conférences contient 2587 références. L'espace des réconciliations conteint donc 6692569 couples de références à comparer. Utiliser l'année comme caractéristique de filtrage, sachant que les publications existent sur 6 années différentes, permet de réduire cet espace de 21,8 %.

L'exploitation de connaissance du domaine telles que les disjonction entre classes (Sais et al. (2007)) : deux références qui appartiennent à des classes disjointes ne sont pas réconciliables. France telecom nous a fournit un corpus décrivant 562368 hotels. Les disjonctions entre classes d'hôtels de pays différents permttent de réduire l'espace des réconciliations de 67,8%.

L'exploitation de propriétés sur les sources telles que l'Unique Name Assumption : deux références issues d'une source de données qui possède la propriété d'UNA sont forcemment distinctes.

Ces techniques de filtrage permettent de filtrer l'espace en déduisant facilement des non réconciliations entre couples. Ce filtrage peut également se propager en cours de traitement. Ainsi, si une référence r1 d'une source S1 à été réconciliée à une référence r2 d'une source S2 qui possède l'UNA, toutes les autres possibilités de réconciliation de r1 avec une référence S2 peuvent être éliminées.

Ces méthodes nous semblent difficilement utilisables lorsque l'on s'intéresse à la réconciliation de schema. Même si deux noms de concepts d'une ontologie correspondent forcemment à des concepts distincts, cette propriété ne peut être propagée dans la mesure où un concept d'une ontologie A peut généraliser deux concepts d'une ontologie B.

2.2 Partitionnement des données pour partitionner l'espace de réconciliation

Le partitionnement consiste à diviser *l'espace de réconciliation* en plusieurs sous-ensembles (parties) de taille plus petite.

Partitions pour la réconciliation : l'espace de réconciliation est partitionné en plusieurs sous-ensembles de paires de références (ou de concepts) de taille plus petites de manière à ce que la couverture (le rappel) de la méthode de réconciliation ne soit pas diminuée.

Pour partitionner l'espace des réconciliations, on peut d'abord partitionner les références (ou les concepts). Pour assurer la non redondance et la non perte d'information, le partionnement doit satisfaire trois critères (Ozsu et Valduriez (1999)) :

Complétude : pour toute référence (resp. concept) il existe une partition P_i contenant cette référence (resp. concept).

Reconstruction : pour toute source S (resp. schéma S) partitionnée en un ensemble de parties P_i, il existe une opération de reconstruction telle que $S = \cup P_i$ pour tout P_i appartenant à l'ensemble des partitions. Cette opération de reconstruction est à définir en fonction du partitionnement effectué. Par exemple, dans le cas où les partitions sont des composantes connexes d'un graphe, l'opération de reconstruction est la fusion de graphes.

Disjonction : une référence (resp. concept) n'est présente que dans une seule partition.

Partitions pour la réconciliation de références : Le graphe G d'un ensemble de références I d'une source de données peut être représenté sous la forme d'un multi-graphe orienté étiqueté dont les sommets V_G sont des références et les arcs E_G sont des relations entre références.

$$G = \langle V_G, E_G, R_G \rangle$$

où

$$V_G = I, \qquad E_G \subseteq V_G \times R_G \times V_G$$

et

$$\langle i1, r, i2 \rangle \in E_G \; ssi \; \exists r, r(i1, i2)$$

Considérons que l'ensemble de références est représenté dans un multi-graphe G. Ainsi, l'ensemble des partitions dans l'ensemble de références correspond exactement à l'ensemble de composantes fortement connexes CFC que l'on peut former dans le graphe G. Une composante connexe pour une référence i représente le graphe contenant l'ensemble des références atteignables à partir de i par les relations instanciées.

Une fois que ces CFC ont été définies, elles sont ensuite enrichies par l'ensemble des valeurs atomiques associées aux différentes références par les attributs. Ces derniers sont très importants pour la mesure de la similarité entre les descriptions des références afin de pouvoir décider de leur réconciliation ou de leur non réconciliation. Ainsi, l'espace de réconciliation de référence est réduit à un ensemble d'espaces plus petits correspondant aux paires de composantes fortement connexes du graphe de références G, enrichies par les valeurs des attributs associés à chaque référence des CFCs. Lors de la réconciliation d'une paire de CFCs, l'ensemble des paires de références est le produits cartésien des deux ensembles de références contenues dans les deux CFCs . Bien sûr, certaines paires pourront ne pas être considérées par la suite compte tenu des connaissances du domaine et des informations renseignées.

Nous notons que la taille des composantes connexes est dépendant du niveau de redondance dans les sources (présence de l'UNA). Plus précisément, plus la source est redondante plus les composantes connexes de cette source sont petites. Ainsi, si une source qui décrit un ensemble de références bibliographiques possède l'UNA, chaque publication est reliée à un ensemble d'auteurs (3 en moyenne) auquels sont également associées d'autres publications qui font donc parties de la même composante connexe. En revanche, une source telle que CORA ne possède par l'UNA et donc chaque composante connexe se limite à la description d'une publication. CORA qui comporte 6000 références peut donc être découpées en 1295 CFC et donc en 1677025 espaces de 25 références en moyenne.

Partitions pour la réconciliation de concepts : Pour la réconciliation de concepts, les concepts de chaque schéma sont généralement organisés dans un graphe connexe où les sommets sont les concepts et où les arcs sont des relations. Dans le cas où il n'existe pas de mappings entre les concepts, comme le graphe des concepts est connexe, on ne peut pas définir de composantes connexes plus petites que le graphe lui-même. En revanche, le partitionnement de l'ensemble des concepts est possible quand des mappings préalables existent dans le cas de taxonomies où la seule relation sémantique considérée est la subsumption. Soient Sc1 et Sc2 deux taxonomies à réconcilier et m un mapping d'équivalence entre les deux concepts A et B calculé ou donné par un expert. Soient G_A et G_B les deux sous-arbres de Sc1 et Sc2 enracinés respectivement par A et par B. L'exploitation du mapping m nous permet de partitioner les deux schemas en quatre sous-arbres :

$$(Sc1 \setminus G_A), (Sc2 \setminus G_B), G_A, G_B$$

Avec ($\setminus$) exprime la difference entre deux graphes.

Ce mode de partitionnement peut être réitéré dans le cas où nous avons à disposition plusieurs mappings. Ainsi, l'espace de réconciliation est décomposé, et donc pour la réconciliation d'une paire de sous-arbre, l'espace de réconciliation est de taille plus petite.

L'ontologie du tourisme fournie par France Telecom comporte 210 concepts. L'espace des réconciliations qui serait créé avec une ontolgie de taille similaire contiendrait donc 44100 couples.

A priori la réconciliation de référence conduit plutôt à de très nombreuses partitions de petites taille tandis que la réconciliation de schema conduit plutôt à un espace de grande taille.

3 Une meilleure gestion du temps

3.1 Approximer

Le traitement de données nombreuses peut conduire à l'utilisation de méthodes permettant d'approximer un résultat grâce à l'utilisation d'algorithmes itératifs que l'on peut décider de stopper après n itérations. De telles méthodes ont pu être définies pour les deux problèmes de réconciliation.

Dans le cas de la réconciliation de schemas, Euzenat et Valtchev (2004) ont défini une méthode itérative utilisant une mesure permettant de comparer les entités (concepts ou relations) de deux ontologies décrites en OWL-Lite. Cette mesure prend en compte toutes les informations concernant un couple de concepts ou de relation (son label, ses propriétés éventuellement multivaluées, ses instances, ses généralisants).

Dans le cas de la réconciliation de référence, nous proposons une mesure de similarité définie pour traiter des données représentées en RDF et décrites par un schema RDFS+ (RDF auquel s'ajoute des opérateurs OWL et des règles SWRL). Nous avons pu prouver qu'un algorithme de calcul itératif utilisant cette mesure converge. Dans ce calcul, la similarité d'une paire de références est fonction de la similarité des paires de références voisines. Les connaissances du domaine sont exploitées afin de faire varier l'impact de la similarité de ces paires voisines, en particulier pour tenir compte du fait qu'une relation est une dépendance fonctionnelle. Ce calcul itératif est appliqué paire de composante connexe par paire de composante connexe.

3.2 Parallélisation de l'exécution de la réconciliation

Afin d'améliorer les performances en terme de temps d'exécution de la réconciliation, une des solutions est de concevoir un algorithme parallèle. Il s'agit de distribuer la tâche de réconciliation sur un ensemble de processus tout en assurant la non perte d'informations. Pour cela, un ensemble de partitions disjointes peuvent être définies comme dans la section 2.

Dans le cas où les partitions ne possèdent pas les propriétés définies dans la section 2(Omar et al. (2006)), propose d'utiliser deux fonctions qui sont importantes pour la distribution des tâches : la fonction "scope" qui permet de distribuer les différentes partitions sur les différents processus et la fonction "responsible" qui permet d'assurer la non redondance, cela en décidant quel processus est responsable de quelle réconciliation. Une définition plus formelle de ces deux fonctions est donnée dans (Omar et al. (2006)).

De la même façon, la parallélisation de la réconciliation d'ontologies est envisageable.

4 Données distribuées, P2P

Les méthodes de réconciliation se confrontent également au fait que les schemas peuvent être distribués. Certaines approches travaillent dans ce cadre et estiment que l'utilisation de mappings entre ontologies locales qui se connaissent est plus réaliste que la création d'une ontologie commune importante. Un système tel que SomeWhere (Adjiman et al. (2006)) est une approche complètement pair à pair dans laquelle chaque pair stocke localement ses propres axiomes et un ensemble de mappings. Il exploite les mappings pour répondre de façon correcte et complète à une requête.

Une telle approche peut être complétée pour permettre de découvrir et d'exploiter des réconciliations de références afin d'intégrer les données provenant de différents pairs lors d'une requête. Nous envisageons une méthode dans laquelle chaque pair connaît éventuellement un ensemble de réconciliations entre les données (références) qu'il possède et celles d'autres pairs (comme pour les mappings). Le traitement des réponses aux requêtes va alors pouvoir exploiter les réconciliations de référence stockées dans les pairs impliqués dans une requête de SomeWhere. L'objectif est de trouver un chemin de reconciliation entre une référence d'un pair et une référence d'un autre pair pour les couples de références pertinentes correspondant à des variables liées de la requête.

5 Conclusion

Nous avons brièvement présenté quelques points de comparaison entre le problème de la réconciliation de références et celui de la réconciliation de schema quand leur résolution se confronte au passage à l'échelle. Ces points de comparaison concernent essentiellement des approches globales, approches qui prennent en compte un nombre d'informations qui peut être important dans un calcul complexe. Différentes stratégies permettant de limiter la taille des données ou le temps de calcul doivent donc être mises en place si ces méthodes veulent passer à l'échelle.

Références

Adjiman, P., P. Chatalic, F. Goasdoué, M.-C. Rousset, et L. Simon (2006). Distributed reasoning in a peer-to-peer setting : Application to the semantic web. *Journal of Artificial Intelligence Research 25*, 269–314.

Baxter R., Christen P., C. T. (2003). A comparison of fast blocking methods for record linkage. In *ACM workshop on Data cleaning Record Linkage and Object identification*.

Euzenat, J. et P. Valtchev (2004). Similarity-based ontology alignment in owl-lite. In *ECAI*, pp. 333–337.

Kalfoglou, Y. et M. Schorlemmer (2003). Ontology mapping : the state of the art. In *Knowl. Eng. Rev., 18(1) :1–31*.

Newcombe, H. B. et J. M. Kennedy (1962). Record linkage : Making maximum use of the discriminating power of identifying information. pp. 563–567.

Omar, B., G.-M. Hector, K. Hideki, L. Tait, M. David, et T. Sutthipong (2006). D-swoosh : A family of algorithms for generic, distributed entity resolution. In *Technical Report, Stanford InfoLab*.

Ozsu, M. T. et P. Valduriez (1999). *Principles of distributed database systems (2nd ed.)*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

Sais, F., N. Pernelle, et M. Rousset (2007). Approche logique pour la réconciliation de références. In *Extraction et Gestion des Connaissances conference (EGC 2007)*.

Stumme, G. et A. Maedche (2001). Fca-merge : bootom-up merging of ontologies. In *17th IJCAI*, pp. 225–230.

Summary

We have compared the ontology mapping and the reference reconciliation problems when they have to deal with numerous and eventually distributed data.