Passage à l'échelle de la réconciliation de concepts et de la réconciliation de références

Nathalie Pernelle et Fatiha Saïs

LRI, Université Paris-Sud XI, INRIA Futurs, Orsay.

Projet PICSEL 3 en collaboration avec France Telecom R&D.

Plan

Motivations

- Diminuer la taille de l'espace de réconciliation
  - Filtrage
  - Partitionnement des données
  - Partitions pour la réconciliation de concepts
- Améliorer les performances en temps
  - Approximation
  - Parallélisation de l'exécution de la réconciliation
- Données distribuées P2P

Conclusion

Motivations : hétérogénéité

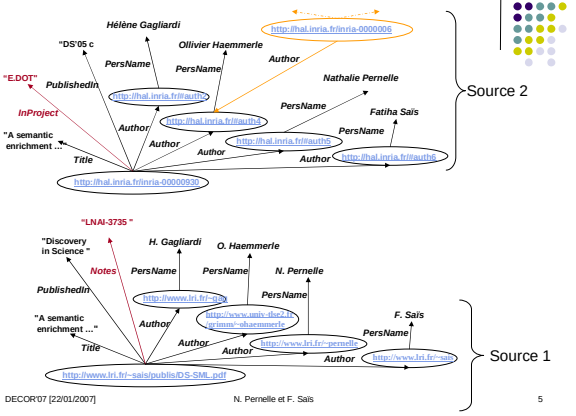
- Un système est **homogène** si le logiciel qui **manipule les données** est identique pour toutes les sources et si toutes les données sont **représentées** selon le **même modèle** (schéma)
- Système hétérogène :
  - **(I) Hétérogénéité des schémas** : représentations différentes
    - Ex: S1 : Personne(NomPers, PrenomPers, Adresse)
      S2 : Individu (Prenom, Nom, Rue, CP, Ville, Fonction)
  - **(II) Hétérogénéité des données** : codage et référentiel différents
    - Ex: S1 : Peinture(*La joconde*) ; S2 : Peinture(*Monalisa*)
      S1 : Musée(*Orsay*) ; S2 : Ville(*Orsay*)
- Une multitude de méthodes de réconciliation de concepts (I) et de réconciliation références (II).

Motivations : Passage à l'échelle

- Confrontation des méthodes de réconciliation à des grandes quantité de données.
  - Ex : sites comparateurs de prix (e.g. www.kelkoo.com) traitent des millions de références de produits pars jours !
  - La quantité de donnée est particulièrement élevée dans les approches globales :

Si 2 nœuds de 2 schémas/ontologies sont similaires, leurs voisins pourraient bien être similaires.
Si deux références sont similaires , leurs références voisines pourraient bien être similaires.

- L'exécution en ligne de la réconciliation la confronte à des contraintes de temps qui peuvent être très strictes (e.g. exécuter 30 requêtes en moins de 3 secondes).
- Un fort besoin de méthodes de réconciliation de références et de concepts passant à l'échelle.



Taille de l'espace de réconciliation : Filtrage

- **Réconciliation de référence** :
  espace de réconciliation = ensemble des couples de références candidats à la réconciliation
- **Réconciliation de schéma,**
  espace de réconciliation = ensemble des couples de concepts et de relations des schémas (ontologies) à réconcilier.
- 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 à comparer.

Taille de l'espace de réconciliations : Filtrage

- **Méthodes de *blocking***
  - On ne considère que les paires de références qui ont une (ou plusieurs) **caractéristiques communes** (Newcombe et al(1962), Baxter R. (2003)).

  Ex : numéro d' ISBN, nom de famille

  - Espace des réconciliations divisé au mieux par m (si m valeurs).

  Ex : Corpus CORA (benchmark de citations), 6000 références (articles, conférences, journaux et auteurs).
  2587 références d'articles, conférences et journaux : 6692569 couples à comparer.
  Utilisation de l'année pour le filtrage (6 valeurs)
  → réduction de l'espace de 21,8 %.

Taille de l'espace de réconciliations : Filtrage

- **Exploitation de la disjonction entre classes**
  Deux références appartenant à des classes disjointes ne sont pas réconciliables.

  Ex : corpus de 562368 hôtels (France telecom).
  Utilisation de disjonctions entre classes d'hôtels de pays différents
  → réduction de l'espace de 67,8%
- **Exploitation de propriétés sur les sources**
  Deux références issues d'une source de données qui possède la propriété d'UNA sont forcément distinctes.
- Méthodes difficilement utilisables lorsque l'on s'intéresse à la réconciliation de schema !

Taille de l'espace de réconciliations : partitionnement des données (1)

- Exploiter la structure des données pour créer des partitions de références et de concepts.
- Respect des contraintes de non redondance et de la non perte d'information [Ozsu et Valduriez (1999)] :
  - Complétude : pour toute référence (resp. concept) il existe une partition contenant cette référence (resp. concept)
  - Reconstruction : pour toute source (resp. schéma) S partitionnée en un ensemble de partitions P, il existe une opération de reconstruction telle que S = (Φ pi), avec pi ∈ P.
  - Disjonction : une référence (resp. un concept) n'est présent que dans une seule partition à la fois.

Taille de l'espace de réconciliations : partitionnement des données (2)

- **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.

$$G = <V_G, E_G, R_G>$$

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

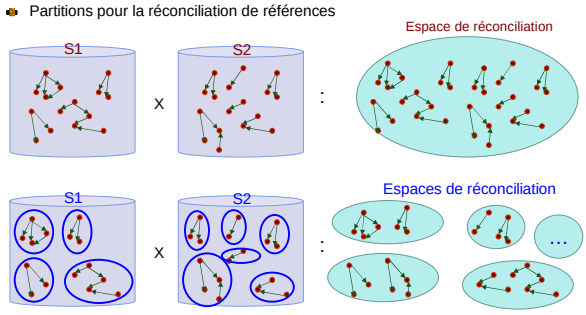
Et $<i1, r, i2> \in E_G$ ssi. $\exists r, r(i1, i2)$

- Former les composantes fortement connexes des graphes de références
- Enrichissement de ces composantes par les valeurs atomiques qui leur sont associées par les attributs. Former les paires de composantes connexes.

  Ex : CORA , espace de réconciliation de 6000x6000 paires de références partitionné en 1677025 espaces de 25 références en moyenne.

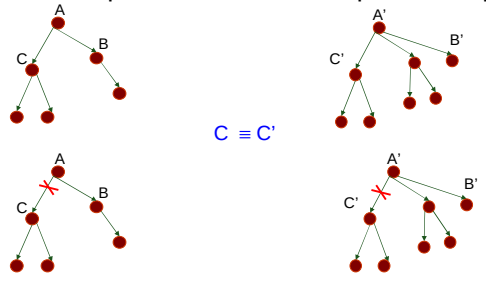
La taille de l'espace de réconciliations : partitionnement des données (3)

- Partitions pour la réconciliation de références



La taille de l'espace de réconciliations : partitionnement des données (4)

- **Partitions pour la réconciliation de concepts. Connexe a priori.**



La taille de l'espace de réconciliations : partitionnement des données (5)

Ex : Une ontologie du tourisme fournie par France Telecom comporte 210 concepts. L'espace des réconciliations créé avec une ontologie de taille similaire contiendrait 44100 couples.

- **Taille des Partitions :**
  La réconciliation de référence conduit plutôt à de très nombreuses partitions de petites taille tandis que la réconciliation de schéma conduit plutôt à un espace de grande taille.

Gestion du temps : Approximation

- Approximer un résultat grâce à l'utilisation d'algorithmes itératifs (s'arrêter après n itérations).
- **Cas de la réconciliation de schemas** :

  OLA: méthode itérative utilisant une mesure permettant de comparer les entités (concepts ou relations) de deux ontologies décrites en OWL-Lite.
  Informations prises en compte : label, propriétés éventuellement multivaluées, instances, généralisants.

Gestion du temps : Approximation

- **Cas de la réconciliation de référence :**
  - Mesure de similarité définie sur des données RDF et décrites par un schema RDFS+ (RDF, opérateurs OWL et règles SWRL).
  - Preuve de convergence d'un algorithme de calcul itératif utilisant cette mesure.
  - Connaissances du domaine exploitées afin de faire varier l'impact de la similarité de paires de références voisines.
  - Calcul itératif appliqué paire de composantes connexes par paire de composantes connexes.

Gestion du temps : méthodes partielles

- Méthodes permettant d'obtenir un **résultat partiel**.

  Eventuellement conçues comme une succession d'étapes.
  Résultat partiel intéressant = produire d'abord les résultats les plus sûrs.

  Réconciliation de concepts (Reynaud et Safar)
  Réconciliation de référence (Saïs et al 2007)

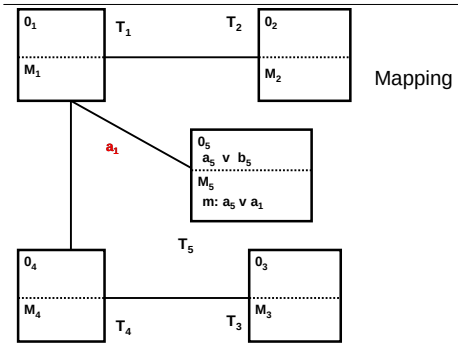
Gestion du temps : parallélisation

- Cas de données partitionnée selon les critères [Ozsu et al] : distribuer l'exécution de la réconciliation sur un ensemble de processeurs en assurant la non perte d'information.
- Dans tous les cas :
  définir des fonctions « scope » et «responsible » [Benjelloun et al.] permettant de distribuer et de vérifier la non redondance des comparaisons.

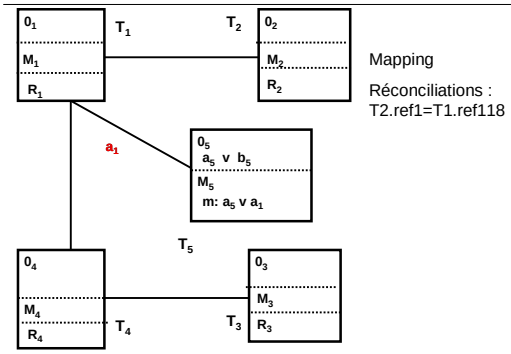
Données distribuées : P2P

- Les schémas/ontologies peuvent être distribués.
- Certaines approches utilisent des mappings entre ontologies locales.
- **SomeWhere** (Adjiman et al. (2006)): approche 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 peut-être être complétée pour permettre de découvrir et d'exploiter des réconciliations de références.

Données distribuées, P2P : plateforme SomeWhere



Données distribuées, P2P : plateforme SomeWhere



Conclusion

- Quelques points de comparaison entre la réconciliation de références et la réconciliation de schéma quand leur résolution se confronte au passage à l'échelle.
- Approches qui se ressemblent, qui prennent en compte un nombre d'informations qui peut être important dans un calcul complexe (approches globales).
- Différentes stratégies permettant de limiter la taille des données ou le temps de calcul doivent être mises en place.