

Sujet de Master recherche

Titre : Utilisation de la théorie des ensembles flous dans la découverte de mappings sémantiques approchés entre schémas.

Mots-clés : bases de données, médiation de données, ensembles flous, mapping sémantique

Lieu : LIRMM

Encadrants : Z. Bellahsene (LIRMM), O. Pivert, H. Jaudoin, A. Hadjali (IRISA)

Objectifs du travail :

Le contexte du travail est celui de la médiation de données. On considère un environnement décentralisé (i.e., il n'existe pas de schéma global unique) où l'on dispose de différentes sources de données (relationnelles ou semi-structurées). Afin de résoudre une requête, ce qui nécessite de la propager vers les sources pertinentes, il est nécessaire de disposer de relations sémantiques entre les schémas de sources de données. L'étape initiale du processus, appelée "schema matching", consiste à mettre en correspondance les éléments (attributs) d'un schéma source avec ceux d'un schéma cible. Sur la base des correspondances obtenues, on cherche ensuite à découvrir des "mappings sémantiques" (ou schema mappings) entre sources de données. Ces mappings peuvent prendre les formes suivantes : $Q1(S1) \equiv Q2(S2)$ ou $Q1(S1) \subseteq Q2(S2)$ où $S1$ et $S2$ sont deux schémas, $Q1(S1)$ et $Q2(S2)$ deux requêtes (flexibles ou non) exprimées respectivement sur $S1$ et $S2$. Ces expressions dénotent respectivement l'égalité et l'inclusion entre les résultats de $Q1$ sur $S1$ et de $Q2$ sur $S2$. L'objectif de ce sujet est d'étudier la notion de mappings *approchés*, considérant qu'en général il est peu probable que deux requêtes sur deux sources différentes produisent exactement le même résultat (ou que le résultat de l'une soit exactement inclus dans celui de l'autre). Ce type de mappings permet notamment d'autoriser la présence d'exceptions dans la satisfaction de l'égalité ou de l'inclusion, ce qui augmente les chances d'obtenir des réponses à une requête donnée dans la mesure où cette requête peut être propagée vers plus de sources (la nature approchée du mapping augmente le nombre de sources cibles). Le cadre théorique que nous suggérons est celui de la théorie des ensembles flous, qui fournit un corpus riche d'opérateurs permettant de modéliser et de traiter les exceptions.

Travail à faire :

- i) une étude bibliographique sur les techniques de schema mapping et sur le passage du schema matching au schema mapping;
- ii) ii) une étude sur la façon d'introduire de la flexibilité dans ces mécanismes (relaxation de l'égalité et de l'inclusion) dans les cas de données relationnelles puis semi-structurées,
- iii) iii) une étude portant sur la fermeture transitive des mappings approchés obtenus (dans le cadre classique, par exemple, si on a $V1 \equiv V2$ et $V2 \subseteq V3$ alors on a $V1 \subseteq V3$) et ses propriétés;
- iv) iv) la mise en évidence de quelques pistes sur la façon d'utiliser les mappings sémantiques approchés pour répondre à une requête (reformulation de la requête, extension de la notion de réponse certaine).

Bibliographie :

E. Rahm, P.A. Bernstein, A survey of approaches to automatic schema matching, *VLDB Journal*, vol. 10, pp. 334-350, 2001.

R. Miller, L. Haas, M. Hernandez, Schema mapping as query discovery, *Proc. of the 26th VLDB Conf.*, 2000.

- L. Popa, Y. Velegrakis, R. Miller, M. Hernandez, R. Fagin, Translating web data, *Proc. of the 28th VLDB Conf.*, 2002.
- Y. Velegrakis, R. Miller, L. Popa, Mapping adaptation under evolving schemas, *Proc. of the 29th VLDB Conf.*, 2003.
- P. Bosc, O. Pivert, About approximate inclusion and its axiomatization, *Fuzzy Sets and Systems*, 157 (11), pp. 1438-1454, 2006.