Optimisation et réécriture de requêtes sous constraintes : Graphes, Arbres et mots

Responsable : Sophie Tison

Encadrants : Pierre Bourhis et Sophie Tison

Equipe : Links Inria Lille/LIFL (UMR CNRS/Université Lille 1/Université Lille 3/Inria)

Contexte : Depuis la création des bases de données, il a toujours été possible d'exprimer des contraintes que les données doivent vérifier. Ces contraintes sont surtout utilisées pour exprimer des spécifications sémantiques que les données respectent. Par exemple, il est possible d'exprimer que chaque individu possède un seul numéro de sécurité sociale. En supposant que les données vérifient ces spécificités, il est possible d'utiliser également ces contraintes pour optimiser une requête posée sur ces données.

Avec la multiplication des sources de données sur le Web et les transformations d'une base de données en une autre, il existe des contraintes sous-jacentes vérifiées par les données. Toutefois, si ces contraintes sont très présentes, elles sont souvent peu utilisées pour optimiser les requêtes.

Problématique : Avec la multiplication de sources de données s'est opérée une simplification de la présentation de données et beaucoup de données sont ainsi représentées sous forme de graphes avec arcs étiquetés, une simplification du modèle RDF de diffusion de données sur le web. Les données étant représentées sous graphes et distribuées, il est important de pouvoir de naviguer efficacement dans ces graphes. Pour cela, des langages de requêtes, comme SPARQL, permettent de mettre en relation des nœuds ayant un chemin satisfaisant une propriété exprimée par un langage régulier. Toutefois, ces langages ont un coût d'évaluation plus élevé que les langages de requêtes classiques. Il est cependant possible de reformuler des requêtes complexes en des requêtes plus simples en utilisant les contraintes. Par exemple, considérons les graphes qui vérifient que si il y a un arc étiqueté par a alors il est suivi par un arc étiqueté b et que si un arc est étiqueté par b, alors il est suivi par un arc étiqueté b. La requête déterminée par le langage *abb+* peut être réécrite en la requête équivalente *a*.

Travail demandé : Le but du stage est de continuer dans la lignée des travaux d'optimisation de réécriture de requêtes en supposant que la base de données satisfasse des contraintes. Dans ce stage, l'étudiant se focalisera à des requêtes et des contraintes exprimables par des langages rationnels. Ceci est motivé par la popularité croissante des langages de requêtes pour interroger les graphes se basant sur les langages rationnels. Dans ce cadre, des problèmes de réécriture de requêtes en exploitant les contraintes peuvent se traduire en termes de langages. Par exemple, un premier problème étudié lors de ce stage sera le suivant :

Soient L et L' deux langages rationnels. Le but est de trouver un langage <<minimal>> L'' tel que

$$L \cap L'' = L \cap L'$$

En particulier, il sera intéressant de savoir quand L'' appartient à des classes langages plus simples, par exemple définissables en logique du premier ordre (FO) ou « piece-wise testable ».

Lorsque le résultat sera positif, il sera intéressant de comprendre la complexité de telles questions et de les implémenter.

Bibliographie :

- Serge Abiteboul, Richard Hull, Victor Vianu: Foundations of Databases. Addison-Wesley 1995, ISBN 0-201-53771-0
- Thomas Place, Marc Zeitoun: Separating regular languages with first-order logic. CSL-LICS 2014: 75
- Thomas Place, Lorijn van Rooijen, Marc Zeitoun: Separating Regular Languages by Piecewise Testable and Unambiguous Languages. MFCS 2013: 729-740
- Wouter Gelade, Tomasz Idziaszek, Wim Martens, Frank Neven, Jan Paredaens: Simplifying XML Schema: Single-type approximations of regular tree languages. J. Comput. Syst. Sci. 79(6): 910-936 (2013)
- Katja Losemann, Wim Martens: The complexity of regular expressions and property paths in SPARQL. ACM Trans. Database Syst. 38(4): 24 (2013)
- Diego Calvanese, Giuseppe De Giacomo, Moshe Y. Vardi: Decidable containment of recursive queries. Theor. Comput. Sci. 336(1): 33-56 (2005)