Réponse à des requêtes conjonctives en présence de règles existentielles – décidabilité, complexité et algorithmes

Michaël Thomazo [1]

Accessit du prix de thèse Gilles Kahn 2014

Michaël Thomazo a soutenu sa thèse [2] *en octobre 2013 à l'Université Montpellier 2, sous la direction de Jean-François Baget et Marie-Laure Mugnier, au sein de l'équipe Inria GraphIK. Il est actuellement boursier Alexander von Humboldt à la Technische Universität Dresden, en Allemagne.*

Les données sont omniprésentes dans notre vie : catalogues de produits en ligne, statistiques de criminalité, données médicales... Les exploiter de manière satisfaisante présente de multiples défis : garantie de la confidentialité, passage à l'échelle, interrogation sémantique... Ma thèse a porté sur ce dernier point : comment enrichir l'interrogation des données à l'aide de connaissances générales d'un domaine ? Le cadre est le suivant : des données sont stockées, par exemple dans une base de données relationnelle, en utilisant un certain vocabulaire. Les liens sémantiques entre les différents termes de ce vocabulaire sont précisés à l'aide d'une *ontologie*. Le but est de répondre à une requête exprimée grâce aux termes de l'ontologie.

1. TU Dresden, Germany.
2. Consultable à l'adresse `https://tel.archives-ouvertes.fr/tel-00925722`

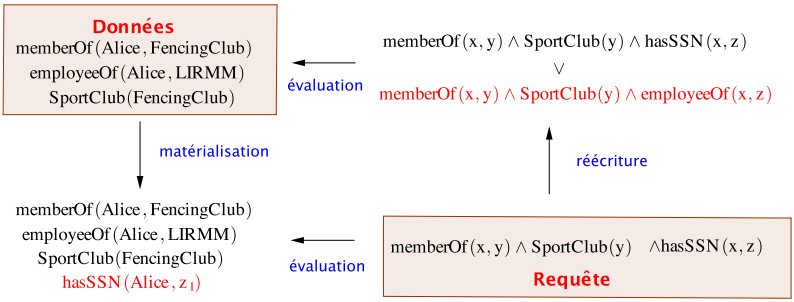

FIGURE 1. Matérialisation et réécriture avec la règle $\forall x \forall y\ (employeeOf(x,y) \rightarrow \exists z\ hasSSN(x,z))$

Le langage de représentation d'ontologie que j'ai étudié durant ma thèse est celui des règles existentielles. Celles-ci possèdent une syntaxe flexible et permettent de décrire des individus qui ne sont pas initialement présents dans la base de données, ce qui est crucial d'un point de vue modélisation. Étant donné des données D, un ensemble de règles existentielles $\mathscr{R}$ et une requête q, le problème est de savoir si q est une conséquence logique de D et de $\mathscr{R}$. Deux grandes approches ont été proposées dans la littérature. La première est une approche de matérialisation, où les règles sont appliquées jusqu'à saturation – on dit que l'on construit le modèle canonique. Une fois ce modèle canonique construit, il suffit d'effectuer une interrogation classique. La deuxième approche est la réécriture de requêtes, où la requête initiale est reformulée en une requête qui est évaluée sur les données initiales. La dualité entre ces deux approches est représentée figure 1, où soit les données sont modifiées (à gauche), soit la requête est modifiée (à droite).

Les procédures résultant de ces deux approches ne sont que des procédures de semi-décidabilité. Dans le cas de la matérialisation par exemple, le modèle canonique peut-être infini. Un important travail de recherche a donc été mené par la communauté pour définir des restrictions sur les ontologies qui assurent la décidabilité du problème. Trois principaux critères sont connus : le modèle canonique est fini, ou il est infini mais de largeur arborescente bornée, ou il existe une réécriture finie en union de requêtes conjonctives.

Dans ma thèse, je me suis intéressé aux deux approches. Premièrement, j'ai étudié les cas où le modèle canonique n'est pas fini, mais où il est possible d'en construire une décomposition arborescente de largeur bornée selon une procédure donnée. J'ai alors montré comment construire une représentation finie du modèle canonique et l'interroger à l'aide d'une opération *ad hoc*, mais de même complexité qu'une simple recherche d'homomorphisme. Cela a fourni le premier algorithme optimal (dans le

pire des cas) pour de nombreuses classes de règles existentielles. Des travaux ultérieurs permettent de réutiliser un moteur Datalog, ouvrant la voie à des implémentations efficaces. De plus, la représentation finie proposée peut servir de point de départ à des développements dans d'autres cadres, en particulier celui du raisonnement en présence d'inconsistances.

Dans un deuxième temps, je me suis intéressé aux approches de réécritures, et en particulier à l'explosion combinatoire qui prend place lorsque de grandes hiérarchies de classes (tout chat est un mammifère, tout mammifère est un animal...) sont présentes dans les ontologies. Des premières expérimentations ont montré que l'approche utilisée, à base de factorisation, permet un gain d'un à deux ordres de grandeur par rapport à l'état de l'art. Je m'intéresse actuellement à des implémentations efficaces des algorithmes présentés dans ma thèse.