Conjunctive query answering under existential rules -Decidability, complexity and algorithms

Michaël Thomazo

La gestion efficace des données et des connaissances est une application importante de l'informatique. Cette gestion fait récemment face à de nouveaux défis, avec notamment la publication massive de données provenant de multiples sources. Cette multiplicité des sources implique une hétérogénéité, qui elle-même rend difficile toute interopérabilité. L'une des solutions largement étudiées, tant au niveau industriel qu'académique, est l'utilisation d'ontologies. Par ontologie, on entend ici une description formelle des connaissances d'un domaine.

L'utilisation de celles-ci présente plusieurs avantages. Premièrement, cela permet de séparer clairement des connaissances générales ("un chat possède quatre pattes"), de celle factuelles ("Tom est un chat"). Deuxièmement, cela permet à un utilisateur d'interroger de multiples bases de données de manière transparente, sans avoir à s'inquiéter des spécificités techniques de chaque source.

Cette thèse se concentre sur un problème intensivement étudié au cours des dernières années : l'évaluation sur une base de données de requêtes conjonctives en présence d'une ontologie. Les requêtes conjonctives sont un choix naturel, dans la mesure où ce sont les requêtes standards en bases de données. Les ontologies sont représentées dans cette thèse à l'aide de règles existentielles, ce qui permet de s'affranchir de certaines limitations imposées par les logiques de description (arité unaire ou binaire, acyclicité), tout en généralisant celles qui sont à la base des standards W3C pour le Web sémantique (notamment, les profils OWL2 QL, RL et EL). Les règles existentielles sont une généralisation de Datalog (langage classique des bases de données déductives), qui a la particularité de pouvoir décrire de nouveaux individus, ce qui est reconnu comme important d'un point de vue modélisation. Cependant, cette capacité rend le problème considéré indécidable.

Deux approches principales ont été utilisées, de manière théorique ou au sein de prototypes, pour résoudre ce problème. La première approche est d'utiliser les règles pour enrichir les données. Le procédé est répété jusqu'à saturation, qui éventuellement ne peut être obtenue qu'après un nombre infini d'étapes. Cependant, lorsque la saturation est obtenue après un nombre fini d'étapes, ou que les faits générés ont une bonne structure (de treewidth bornée), le problème est décidable. La deuxième approche consiste à réécrire la requête initiale en une formule logique du premier ordre à l'aide de l'ontologie. Cette formule est ensuite évaluée directement sur les données, en ne faisant plus attention à l'ontologie. L'idée originale était alors de se servir des systèmes de gestion de bases de données existants pour s'occuper efficacement de cette deuxième partie.

Cette thèse apporte des contributions relatives aux deux approches. Premièrement, aucun algorithme n'était connu pour la plupart des cas pour lesquels aucune des deux approches mentionnées ci-dessus ne terminent. De manière à proposer un algorithme générique, une nouvelle classe décidable a été proposée, généralisant de manière significative les classes existantes. De plus, un algorithme optimal dans le pire des cas pour cette classe de règles a été défini. Cet algorithme peut également être adapté de manière simple pour obtenir des algorithmes optimaux pour chacune des sous-classes connues à ce jour. Il a donc permis d'avoir une vue unifiée et une compréhension

fine de la décidabilité et de la complexité du problème de réponse à des requêtes conjonctives sous une grande variété de classes de règles.

Deuxièmement, la réécriture des requêtes a été étudiée. Au cours des expérimentations menées par la communauté, il est vite apparu que la taille des réécritures obtenues est rédhibitoire, même pour des requêtes et des ontologies simples. En effet, les systèmes proposés au moment de l'écriture de la thèse utilisaient des unions de requêtes conjonctives. Les réécritures typiquement obtenues étaient des unions de plusieurs dizaines de milliers de requêtes conjonctives — au-delà de ce qui peut être traité par les systèmes actuels. Partant du constat que ces importantes tailles étaient dues au choix du langage de réécriture, il est développé dans la thèse un nouvel algorithme de réécriture, dont la spécificité la plus évidente est d'utiliser des formules logiques mieux adaptées aux ontologies existantes. En effet, celles-ci comportent classiquement une importante partie taxonomique. Une utilisation renforcée de la disjonction permet d'éviter l'explosion combinatoire sur les benchmarks usuels. Cet algorithme peut s'utiliser sur une classe de règles bien plus large que la plupart des algorithmes existants, qui se contentent généralement de considérer OWL 2 QL. En plus du design de l'algorithme, le travail de thèse inclut une implémentation, qui est en cours d'intégration à une plate-forme développée depuis mars au sein de l'équipe GraphIK.

De nombreux développements sont en cours d'exploration. Pour commencer, le premier algorithme mentionné est en cours d'implémentation dans le cadre d'une bourse Alexander von Humboldt. Le résultat à venir sera le premier prototype à pouvoir raisonner en présence de règles gardées (une des restrictions les mieux connues, mais uniquement d'un point de vue théorique). Ces règles sont en particulier intéressantes car elles généralisent les logiques de description légères qui sont à la base du Web sémantique. Leur expressivité est cependant bien plus grande, et il est à espérer que ce prototype incitera les applications utilisant cette expressivité. Ce travail fait, il sera intéressant de s'appuyer dessus pour développer des mécanismes de raisonnement approché et de raisonnement en présence d'inconsistance.