

GRANDES BASES DE CONNAISSANCES : STOCKAGE ET DISTRIBUTION

J. F. Baget – baget@lirmm.fr

M. Croitoru – croitoru@lirmm.fr

CONTEXTE GENERAL

Nos travaux en représentation de connaissances portent sur des bases de connaissances dont les éléments sont de 2 types principaux :

- Les *faits* (« Garfield est un chat orange il est assis sur un canapé rouge ») sont exprimables dans le fragment positif, conjonctif, existentiel de la logique du 1^{er} ordre
- Les *règles* (« Tous les chats sont des animaux à 4 pattes, ayant une queue ») encodent des connaissances universelles, leur expressivité est celle de contraintes d'intégrité très générales en bases de données, appelées TGDs (« tuple generating dependencies »).

Nous étudions des algorithmes efficaces permettant, étant donnée une telle base de connaissances, de décider si un fait particulier (la question « Y-a-t'il un animal ayant une queue sur un canapé ? ») est déductible de la base.

Dans le cas où la base est réduite à un ensemble de faits, la déduction est un problème NP-complet, et les optimisations algorithmiques sont similaires à celles utilisées pour optimiser le backtrack dans les réseaux de contraintes.

Lorsque la base comprend également des règles, la déduction est un problème indécidable. Deux types d'algorithme sont utilisés :

- Le chaînage avant enrichit la base de faits avec tout ce que les règles permettent de déduire.
- Le chaînage arrière réécrit la question à l'aide des règles jusqu'à obtenir une nouvelle question à laquelle la base de faits permet de répondre.

Des travaux plus récents nous ont permis d'isoler des sous-classes décidables du problème de déduction combinant ces 2 algorithmes.

VERS DE GRANDES BASES DE FAITS DISTRIBUEES.

L'application de ces travaux théoriques à de « vraies » applications gérant de grandes masses de données (Web Sémantique, bases de connaissances scientifiques en agronomie, bases de données bibliographiques, etc.) soulève de nouveaux défis. Nous devons alors considérer :

1. Des bases de faits de très grande taille, ce qui pose des problèmes de stockage et d'accès.
2. Des bases de faits distribuées (soit intrinsèquement, dans le cas du web, soit parce qu'il n'est pas possible de stocker / charger une grande base).
3. Des bases de faits stockées suivant des paradigmes différents (par exemple, des graphes dans le cas de graphes conceptuels, des tables dans le cas des bases de données relationnelles, ou un ensemble de triplets dans certains « triple stores » RDF).

AXES DE RECHERCHE

Ces nouveaux défis ouvrent de nombreuses pistes de recherche :

1. Etude comparative de différents types de stockage, de leur efficacité respective suivant les caractéristiques de la base de faits encodée (temps d'accès pour les opérations élémentaires), et utilisation de ces caractéristiques pour des optimisations dédiées au backtrack.
2. Distribution du mécanisme de backtrack pour interroger simultanément de nombreuses bases existantes (indexation, réécriture de la requête, agrégation de résultats partiels).
3. (nécessite 1+2) Etant donnée une gigantesque base de faits, comment la découper de façon à accéder à ses parties le plus efficacement possible ?

Au vu de la multiplicité des axes de recherche proposés, l'étudiant pourra se focaliser en fonction de ses intérêts et compétences sur:

- Un travail théorique sur la distribution et le découpage de grands graphes
- Un travail plus applicatif sur une évaluation des différentes techniques de stockage et d'indexation de bases de faits (notamment de graphes).