LARGE GRAPH KNOWLEDGE BASE: STORAGE AND DISTRIBUTION

J. F. Baget - baget@lirmm.fr
M. Croitoru - croitoru@lirmm.fr

GENERAL CONTEXT

Our knowledge representation work focuses on knowledge bases composed of :
- Facts ("Garfield is an orange cat is sitting on a red sofa"). The facts can be represented in the positive, conjunctive, existential fragment of 1st order logic.
- Rules ("All cats are animals to 4 feet, with a tail") encode universal knowledge; their expressiveness corresponds to integrity constraints in databases called TGDs (Tuple Generating Dependencies).

Given knowledge base as described above we study efficient algorithms for deciding whether a particular fact is deductible from the knowledge base (for instance answering the question "Is there an animal with a tail on a sofa?").
When the knowledge base is reduced to a set of facts, the inference is an NP-complete problem, and its algorithmic optimizations are similar to those used to optimize the backtrack in constraint networks.
When the knowledge base also includes rules, the deduction is an undecidable problem. Two types of algorithm are then used:
- The forward chaining enlarges the basis of facts with all the rules that can be deduced.
- The backward chaining rewrites the query with the rules until a new query on which the basis of facts can be answered.
Recent work has allowed identifying decidable subclasses of the above mentioned problem of by combining forward and backwards chaining.

TOWARDS LARGE DISTRIBUTED KNOWLEDGE BASES

This theoretical work in the context of real world applications (see for instance the Semantic Web, scientific knowledge bases in agronomy, bibliographic knowledge bases, etc..) raises a number of new challenges. We must then consider:
1. Very large knowledge bases, which poses problems of storage and access.
2. Distributed knowledge bases (either intrinsically on web, or since is not possible to store / load large bases).
3. Bases of facts stored according to different paradigms (graphs in the case of conceptual graphs, tables in the case of relational databases, or set of tuples in some "triple-stores" RDF).

RESEARCH THEMES

These new challenges raise many research directions:
> **1.** The comparative study of different types of storage, their respective efficiency depending on the characteristics of the basic facts encoded (access time for elementary operations), and the usage of these features for optimization dedicated to backtrack.
> **2.** The distribution of backtracking mechanism for simultaneous interrogation of many existing databases (indexing, rewriting the query, aggregation of partial results).
> **3.** (Requires 1 +2) Given a huge database of facts how to partition it in order to access its parts as efficiently as possible.

The student can focus depending on their interests and skills on:
- A theoretical study on the distribution and partitions of large graphs
- Application oriented work of different techniques for storing and indexing knowledge bases of facts (including graphs).