Static Analysis and Optimization Questions of Structural Recursions in Semistructured Databases

András Benczúr and Balázs Kósa

Since both semistructured and XML data are modeled with rooted, labeled, directed graphs, each query language concerning such data types should contain operations for traversing and complex restructuring of data graphs. In our paper we examine such an operation, namely structural recursion introduced in UnQL, whose underlying algebra, UnCal, is a conservative extension of relational algebra over relational data. The role of structural recursions in UnCal can be described, as it is similar to the role of SPR queries in SPJRU algebras on relational databases, i.e., the algebras consisting of the operators: selection, projection, join, rename and union. Bunemann et al. offered powerful optimizations, when two structural recursions is to be computed in sequence. Dan Suciu showed its very advantageous characteristics in distributed systems. In addition it forms the core language of XSL, the first commercial XML query language.

The data model of UnQL, like the relational model, is "value based", i.e. object identifiers are not assigned to nodes. Two data graphs is considered equivalent, if they are bisimilar. Semistructured data is often described as "scema-less" or "self-describing", since information, which are part of the schema in traditional databases, is intermingled with data. However, various methods were developed to represent even partially knowledge of the structure of data. One of them, which suits well the data model of UnQL, uses schema graphs and dual schema graphs.

In the course of static analysis certain properties of queries given with their syntax is examined without running them. One possible question is that, for a given query q whether there exists an instance I s.t. $q(I)$ is not empty. We introduce a new semantics for structural recursions, which is equivalent to that of defined earlier. With the aid of this the previous question can be answered easily in linear time. Partly owing to its relation with optimization questions mentioned formerly, we also examine the above question in that case, when the inputs and outputs of q are restricted by means of schema graphs and dual schema graphs respectively. The auxiliary graphs introduced in the new semantics is turn out to be a very useful tool for representing relationships among schema and data graphs. Our algorithms also give us other optimization methods. The optimization of regular path expressions in the presence of schema graphs was studied earlier. Regular path expressions can be encoded as structural recursions and one of our method in this paper is a generalization of a former result. Others are different from it, but in the background we always use the same technique given by the new semantics.

The usefulness of the new semantics becomes more obvious, when structural recursion with conditions introduced in UnCal are to considered. By means of this the complex relationships among conditions can be described and the unnecessary ones can be recognized. However, the former questions are NP complete in this case. The discussion is presented in another paper.