Data Integration under the Schema Tuple Query Assumption

Michael J. Minock

Department of Computing Science,
The University of Umea,
Umea, Sweden
mjm@cs.unu.se

Typically data integration systems have significant gaps of coverage over the global (or mediated) schema they purport to cover. Given this reality, users are interested in knowing exactly which part of their query is supported by the available data sources. This work introduces a set of assumptions which enable users to obtain intensional descriptions of the certain, uncertain and missing answers to their queries given the available data sources. The general assumption is that query and source descriptions are written as tuple relational queries which return only whole schema tuples as answers. More specifically, queries and source descriptions must be within an identified sub-class of these 'schema tuple queries' which is closed over syntactic query difference. Because this identified query class is decidable for satisfiability, query containment and equivalence are also decidable. Sidestepping the schema tuple query assumption, the identified query class is more expressive than conjunctive queries with negated subgoals. The ability to directly express members of the query class in standard SQL makes this work immediately applicable in a wide variety of contexts.