Project description: Cost-based Data Integration with Disjunction

Michael Benedikt

1 Project Background

Data integration take as input some meta-data about a querying scenario – e.g. descriptions of sources and mappings from sources to a global schema) along with a user query, written over a global schema. They then answer the query by issuing queries to the sources and merging the results. The PDQ system [3, 1, 2] supports cost-based data integration: the system has additional information concerning the cost of input operations, and looks not just for any means to answer the query, but the lowest-cost one. PDQ also incorporates reasoning about sources and their relationships to global schema tables and to each other, given as dependencies.

PDQ can currently only simple relationships between global schema relations and local sources. To support more realistic relationships requires support for *union* at the plan level. This in turn requires supporting *disjunction* in the description of relationships of sources to global schema tables: it is critical to support global schema relations that are a disjunction of local sources. The goal of the project would be to investigate how to support this; it would be a mix of algorithm development, implementation and experiment on top of the PDQ, and investigation of new reasoning techniques, with the exact balance determined by the fit for the student. For algorithm development, a main issue is to revise PDQs search algorithm to handle disjunction. For implementation and experiment, we have some benchmark scenarios from biology (taken from the European Bioinformatics Institute (see [4]), and would like to test out methods for data integration on them. For reasoning techniques, a key goal is to support resolution-based theorem-proving within PDQ, since this is a good fit for disjunctive reasoning. We have some preliminary results on resolution, but a number of practical and theoretical issues still require investigation.

References

[1] M. Benedikt, J. Leblay, and E. Tsamoura. PDQ: Proof-driven query answering over web-based data. In *VLDB*, 2014.

[2] M. Benedikt, J. Leblay, and E. Tsamoura. Querying with access patterns and integrity constraints. In *VLDB*, 2015.

[3] Michael Benedikt, Balden Ten Cate, Julien Leblay, and Efthymia Tsamoura. *Generating plans from proofs: the interpolation-based approach to query reformulation*. Morgan Claypool, 2016.

[4] Michael Benedikt, Rodrigo Lopez-Serrano, and Efthymia Tsamoura. Biological web services: Integration, optimization, and reasoning. In *Workshop on Advances in Bioinformatics and Artificial Intelligence: Bridging the Gap*, pages 21–27, 2016.