KeywDB: Keyword Driven Mapping Construction

D. Zheleznyakov, E. Kharlamov, I. Horrocks
V. Klungre, M. G. Skjæveland, D. Hovland, M. Giese, A. Waaler

University of Oxford
University of Oslo

Semantic Access to Databases

- Optimised for query answering
- Historically evolve in user-unfriendly form
- Statoil
  - Exploration & Production Data Store (EPDS)
  - Has been developed for 15 years
  - 3K tables, 37K columns, 700 GB data

Ontology Based Data Access

- Ontology: conceptual domain model
- Mappings: relate ontological terms to DBs

Connecting Data to Ontologies

- Problems
  - Connect new DBs to the ontology
  - Add new vocabulary to the Ontology

- Existing approaches
  - Direct mappings: mirror the structure
  - May not work in many applications

- Project Goals:
  - Facilitate discovery of mappings that reflect users’ expectations
  - Enable discovering of quality mappings in industry: Statoil

Keyword Driven Approach: General Idea

User

- Describes: what she expects from the ontology
- Provides: examples of entities of the missing class C
- Example = set of keywords
- Keyword = a characteristics, or attribute value for entities in C

System

- Returns a ranked list of queries
  - $f_1$: SQL, $f_2$: SQL, ..., $f_n$: SQL
- Each query represents C
- In materialisation of each query SQL $q_i$
  - each tuple corresponds to an entity of C
  - some user’s ex. are “among” the tuples
- The higher the rank, the better
  - the query captures user’s expectations

Research Challenges

- Data graph: too large – good to define semantics
- Schema graph: does not help much (no keyword info)
- Keyword driven schema graph: good balance, practical

Graphs

- Candidate sub-graph selection
- Indexes for keyword match, node reachability
- Effectiveness
  - Target queries are in top-k
  - Small number of “simple” keywords is enough
- Top-K queries
  - Top-k without exact ranking
  - Approximation of ranking

KeywDB System

- Main features
  - Allows for multiple examples, each with several keywords
  - Computation of
    - Schema graph, keywords driven schema graph
    - Inverted index for keywords
    - Reachability index for keywords driven schema graph
  - Support for mapping configuration via attribute selection

- Flexible configuration
  - Top-k, maximal query size, query similarity

Optique