Abstract. We have built a virtual data integration system which extends the BIRN Mediator[1] to include OWL 2 ontologies where these ontologies are used as domain schema. This mediator integrates data from multiple, heterogeneous sources. This data integration system rewrites conjunctive queries using LAV views defined over a DL-Lite ontology. This is an important problem in query optimization, data integration and ontology-based data access. The mediator presents to users a unified domain ontology defined in DL-Lite. However, the mediator contains no data, all the data remains at the sources. Each source may have a different schema. We define LAV logical mappings from each source to the mediator's domain ontology. Given a conjunctive user query, our mediator first rewrites it into a union of conjunctive queries (UCQ), which captures the inferences implied by the DL-Lite ontology, using REQUIEM[2]. Second, the mediator rewrites this domain UCQ, using the source mappings, into a UCQ over the source schemas. Third, this UCQ over the source schemas is converted into a SQL query and is evaluated via OGSA-DAI resources and is sent to the MySQL database via the query execution engine.

1. Introduction

OWL 2 QL is based on DL-Lite(r) -- one of a family of description logics developed by Calvanese et al[3]. The DL-Lite(r) rewriting algorithm of Calvanese et al. [3], transforms a conjunctive query Q and a DL-Lite(r) ontology O into a union of conjunctive queries Q1 such that the answers to Q on any set of instance data A can be obtained by evaluating Q1 over A only.

The Mediator system that we have built is divided into three phases. In the first phase we use REQUIEM which takes as input a conjunctive query Q and the ontology O, and uses a resolution-based technique to produce a rewritten query Q1 which is a union of conjunctive queries.

Suppose we have a table Professor with attributes name, department, and telephone in the domain schema; and a table Student with attributes name, major, address, and tutor. Suppose we use the following ontology O[2] to describe the domain schema:

\[
\text{Teacher} \sqsubseteq \exists \text{teaches} \tag{1} \\
\text{Professor} \sqsubseteq \text{Teacher} \tag{2} \\
\exists \text{hasTutor} \sqsubseteq \text{Professor} \tag{3}
\]
Axiom (1) states that teachers teach at least someone, axiom (2) states that professors are teachers, and axiom (3) states that the range of the property hasTutor is Professor. The ontology is used to rewrite the query into a union of conjunctive queries. For example, consider the query

\[ Q_0(x) \leftarrow \text{teaches}(x,y) \]  

posed over O. The rewriting Q1 of query (4) w.r.t. O contains (4) and the following queries:

\[ Q_0(x) \leftarrow \text{Teacher}(x) \]  

\[ Q_0(x) \leftarrow \text{Professor}(x) \]  

\[ Q_0(x) \leftarrow \text{hasTutor}(y,x) \]

Intuitively if asking for all individuals that treat some patient, and given the incompleteness of the sources in the data integration context, we should also query the sources for all Teachers (since all teachers teach students by axiom (1), and hence all Professors and all the range of hasTutor relation.

Now we have ontology-expanded queries which use predicates from the domain schema, and the next phase will reformulate them using the LAV rules. The second phase of our system is answering Q1 using Scalable graph based query rewriting using views or source descriptions(GQR)[4]. GQR looks at the problem from a graph perspective and compactly represents common patterns in the source descriptions. It rewrites a conjunctive query using LAV descriptions, to a maximally-contained union of conjunctive rewritings. This gives union of conjunctive rewritings over the source schema in Datalog format which is converted into a SQL query.

This SQL query is further evaluated with the OGSA-DAI Distributed Query Processing resource and is then sent to the MySQL data resource. OGSA-DAI is a framework that executes workflows which contain activities. Each activity is a well-defined functional unit - data goes in, something is done, data comes out.

2. Approach and Data Sources

The architecture of our virtual data integration system is described in Fig. 1. We have created synthetic source descriptions based on the requiem expansion of the input query. We have created 50 sources for every ontology and a LAV mapping for every source. These sources are MySQL database tables. We have used ontologies as the domain model. We have experimented with 4 ontologies which we have taken from REQUIEM[2]. An example of a LAV mapping would look like:

\[ S4(a0,a1) \rightarrow \text{Course}(a2)^\text{advisor}(a0,a1)^\text{takesCourse}(a0,a2)^\text{teacherOf}(a1,a2)^\text{UndergraduateStudent}(a3) \]

where S4 is an actual table in a MySQL database and the predicates in the body of this LAV mapping are predicates defined in the domain schema which in our case is an ontology.
Fig. 1: Virtual data integration system, BIRN using ontologies as the domain schema

We take ontology file, the LAV rules and the Input Query as inputs to this virtual integration system. Ontology is used as the mediator schema or domain schema. REQUIEM takes this input query and converts it into a union of conjunctive queries using a resolution based technique. Now graph based query rewriting (GQR) reads the source descriptions or LAV rules and reformulates the REQUIEM output which is in terms of mediator predicates and is a union of conjunctive queries in terms of source schema. This union of queries is converted into a SQL query and is finally sent to the Query execution engine which sends it to the MySQL data sources.

3. Evaluation and Challenges

We have done some experiments on this new BIRN data integrating system with four ontologies. These ontologies are vicodi, stockexchange, university and adolena. We have taken these ontologies from REQUIEM[2]. We have created 50 data sources for every ontology and a LAV mapping to domain schema predicates for every source. We have calculated Requiem expansion size and GQR rewritings size for 25 as well as 50 source descriptions, the results are in Fig. 3 and 2 respectively. We have also calculate Requiem expansion time, GQR rewriting time and the preprocessing time, which is the time taken by GQR to read the source descriptions into data structures. We have done this too for both 25 and 50 descriptions, the results are in Fig. 5 and Fig. 4 respectively.
Fig. 2 Requiem Expansion size and GQR Rewritings size for 50 source descriptions. (logarithmic scale)

Fig. 3 Requiem Expansion size and GQR Rewritings size for 25 source descriptions. (logarithmic scale)
One thing to notice from Fig. 2 and Fig. 3 is that GQR rewritings produced for stockexchange ontology are of big size (and this graph is in logarithmic scale). The rewritings are in size of millions, one reason behind this is that the requiem expansion of stockexchange ontology had very less unique predicates and most of them were repeated a lot of times which made it harder to model the source descriptions as unique. One of the challenges in this project is putting together the pieces of the BIRN Mediator integration system, which includes using ontology as the domain schema, integrating REQUIEM and using LAV source descriptions for GQR rewritings. The
other challenge as I have mentioned above is to model the source descriptions to control the size of GQR rewritings. Since rewritings evaluation by RDBMSs may be costly or even unfeasible, trying to produce small rewritings is of critical importance.

4. Conclusion

We now have a virtual data integration system which extends the BIRN Mediator[1] to include OWL 2 ontologies and it also supports LAV rewritings using views. Now we can deal with a set of constraints(ontology) on the domain schema which lie on top of the mediator. We have also experimented with synthetic MySQL data sources to measure Requiem expansion size, GQR rewritings size, Requiem Expansion time, preprocessing time of source descriptions and GQR Rewriting time. The challenging task was to make this as an end to end system by integrating small parts together. Source modeling was also challenging because we wanted to control the reformulation size of GQR. As a future work, we will further extend the BIRN Mediator to include GLAV(Global – Local as View) rewriting as well. We will also extend the BIRN Mediator by integrating Karma[5] into it. Karma reads real data sources as well as an ontology to give source descriptions which our current system can use to rewrite an input query to a union of conjunctive queries.

5. References

1. Naveen Ashish, Jose Luis Ambite, Maria Muslea and Jessica A. Turner, Neuroscience Data Integration through Mediation: An (F)BIRN Case Study. Frontiers in neuroinformatics, 2010. 4:118


