Querying the Web On-the-fly Using Ontologies and Mappings

Konstantina Bereta, George Papadakis, and Manolis Koubarakis

National and Kapodistrian University of Athens
{Konstantina.Bereta, gpapadis, koubarak}@di.uoa.gr

Despite the rapidly increasing availability of open data as Linked Data on the Web, a lot of data is still published in other, non-RDF formats, such as HTML forms/tables, or via Rest APIs. To use this data as linked data, one would have to transform it into RDF triples and store them in a triple store every time a source gets updated. The first OBDA/ RDB2RDF systems started to emerge recently, providing a solution for creating virtual RDF graphs on top of relational data using ontologies and mappings, such as Ontop [1], Ultrawrap [2] and Sparqlify[1].

In this work, we introduce a framework for extending existing OBDA techniques to support querying of data from different sources that are available on the Web, such as webtables and Rest APIs.

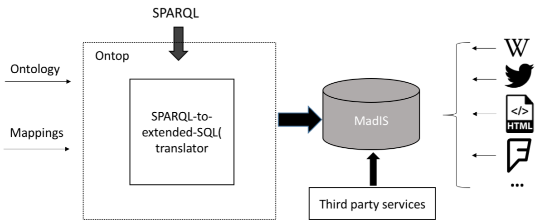

Fig. 1: Ontop4theWeb architecture.

Our approach is implemented in the system Ontop4theWeb and Figure 1 shows its architecture. The following steps briefly describe our approach.

First, we extend SQL with virtual table operators. We implement a virtual table operator for each data source that we want to access and this operator is used embedded in SQL queries; when invoked, the operator creates a virtual table on-the-fly, populating it with the data currently retrieved from the selected data source. Thus, the retrieved data can now be handled as relational.

We allow for SQL-extended queries described above in R2RML mappings (instead of standard SQL ones). The mappings encode how the relational data can be mapped into RDF terms. An example is provided below.

[1] http://aksw.org/Projects/Sparqlify.html

```
mappingId webtable_rotten_tomatoes
target rot:{rank} rot:rank {rank}; rot:title {Title} .
source select  rid as rank, Title from
       webtable(’http://www.rottentomatoes.com/ top/bestofrt/’, 3)
```

The mapping described above encodes how the attributes of the resulting virtual table (i.e., *title* and *rank* in our example) are mapped into RDF terms. In this way, users are able to pose queries that retrieve the rank and the title of rotten tomatoes movies by querying the respective HTML table directly, using SPARQL.

```
select distinct  ?title
where { ?s rot:title ?title . ?s2 rot:rank ?rank }
```

When a SPARQL query as the one described above is posed, it gets translated into the respective SQL query that contains the virtual table operator that corresponds to the data source (i.e., webtable). The operator populates the virtual table with data and after the whole SQL query is evaluated the results are returned as virtual RDF terms to the users.

To improve performance, we allow the option to use cached data for a time window w. If w is expired, fresh data needs to be retrieved. The length of w is a parameter that can be configured per mapping (i.e, different mappings can access the same data source at different rates depending on the needs).

Ontop4theWeb is based on the system Ontop[2] [1], a state-of-the-art, open-source OBDA system that supports both R2RML and the OBDA mapping language. More specifically, it extends the geospatial extension of Ontop named Ontop-spatial[3] [3]. As back-end, it uses the MadIS[4] [4] system, an extensible relational database system built on top of the SQLite[5] database that enables users to implement user-defined functions as virtual table operators.

We have implemented virtual operators that access Twitter and Foursquare APIs, and we have extended the MadIS webtable operator that retrieves data from HTML tables.

To evaluate our approach, we conducted experiments that involved SPARQL queries that access different sources (Twitter API, Webtables) as well as generated datasets (i.e, generated webtables to measure scalability) and we measured the query execution time. Even in cases where no cache is used, Ontop4theWeb can execute queries in a few seconds, whereas when the cache is used, queries are executed in less than a second. The experiments also showed that Ontop4theWeb is able to query webtables with up to 100,000 rows within minutes.

References

[1] Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., Xiao, G.: Ontop: Answering SPARQL queries over relational databases. Semantic Web **8**(3) (2017) 471–487

[2] https://github.com/ontop/ontop

[3] http://ontop-spatial.di.uoa.gr/

[4] http://madgik.github.io/madis

[5] http://www.sqlite.org

[2] Sequeda, J.F., Miranker, D.P.: Ultrawrap: SPARQL execution on relational data. J. Web Sem. **22** (2013) 19–39

[3] Bereta, K., Koubarakis, M.: Ontop of Geospatial Databases. In: Proceedings of the 15th International Semantic Web Conference. (2016)

[4] Chronis, Y., Foufoulas, Y., Nikolopoulos, V., et al: A Relational Approach to Complex Dataflows. In: EDBT/ICDT Workshops. (2016)