Incremental SPARQL query processing

Ana I. Torre-Bastida

Tecnalia Research & Innovation
Parque Tecnológico Edif 202
Zamudio, 48170 - Vizcaya
isabel.torre@tecnalia.com

Abstract. The number of linked data sources available on the Web is growing at a rapid rate. Moreover, users are showing an interest for any framework that allows them to obtain answers, for a formulated query, accessing heterogeneous data sources without the need of explicitly specifying the sources to answer the query. Our proposal focus on that interest and its goal is to build a system capable of answering to user queries in an incremental way. Each time a different data source is accessed the previous answer is eventually enriched. Brokering across the data sources is enabled by using source mapping relationships. User queries are rewritten using those mappings in order to obtain translations of the original query across data sources. Semantically equivalent translations are first looked for, but semantically approximated ones are generated if equivalence is not achieved. Well defined metrics are considered to estimate the information loss, if any.

Keywords: Semantic Web, Linked Open Data Sources, query reformulation, query rewriting, ontology mapping.

1 Problem statement and research question

The Linked Open Data (LOD) initiative has made available to the users a large number of data sources from various domains such as education, life sciences, government data, literature, geography and others. Two commonly used approaches for query processing in this context are: 1) to query the different data sources independently, one by one; or 2) to integrate first the data sources into a local centralized warehouse and then to process queries in a centralized way on the warehouse. Both approaches present relevant problems such as the user needed expertise following the first approach and the scalability problems that arise in the second one. In this scenario an alternative approach is appearing, the so called federated approach, in which a query is formulated and its answer is obtained from different sources but with the distinguishing feature that the technical details associated to the distributed query answering process are transparent to the user. The work developed in this thesis is placed in this approach, but our system will have the added feature that the user does not need to have specific knowledge of the language in whichthe different data sources are modeled. We summarize our research question as the following one: *How*

can we assist to the user with querying heterogeneous data sources, without the need to be an expert on the ontologies with which they are modeled and returning incremental and satisfactory results for the user?

Consider the following scenario, a music student formulates the following query to a multimedia local source: recording of "Sonata giocosa" by "J. Rodrigo" played by "Marco Socias". Possibly, due to the limited number of records of that source, the answer to that query is empty. Then, the user clicks on not satisfied and requests the system to reformulate the question about the same source. Transparently to the user, the system reconstructs the query as: recording of "Sonata giocosa" by "J. Rodrigo" played by anybody. This time the answer received is a recording of the requested piece from the polish guitarist "Marcin Dylla". The user clicks again on not satisfied and on this occasion asks the system to consult a new source. The system selects another relevant source or the user can select the source from a set provided by the system. In this case, the user leaves the decision in the hands of the system and it choices a source consisting of cd records and reformulates the query as: cd including "Sonata giocosa" and featuring "Marco Socias". This time the answer is the cd with title "Elogio de la guitarra" where "Marco Socias" plays the requested piece. Notice that in those last cases the semantics of the original query has been changed.

According to my proposal, the user formulates a query expressed with her preferred vocabulary, waits for an answer and asks for more answers if she is not satisfied with those received. Then, the system does its best to satisfy the user. If semantically equivalent translations of the original query are not achievable on different sources, the system proceeds with approximate translations with the hope to find satisfying answers for the user. The system is able to measure the incurred loss of information with the approximate translation, using metrics from the field of information retrieval, such as precision and recall.

The novel contribution that I consider is: *An innovative query approach that provides the answers by accessing different data sources, expressed with different vocabularies, in an incremental way guided by the user. Source mappings are used for issuing translations of the original query and a measure of loss of information incurred in the intended translation is provided in the case that it occurs.*

2 State of the art

The Sparql query processing over heterogeneous data sources is an extensive research field in the Semantic Web community. Currently, many systems (DarQ[6], FedX[8]) deal with query federation on heterogeneous datasources of the Web of Data[1]. But the federated approach has a fundamental difference with ours, this is the need for the users to know the ontologies with which are described the datasets and write the query in their model. Our approach is more flexible and useful to the user who only knows his dataset domain and languages, being the system responsible of rewrite the query in terms of the ontologies of other additional interesting datasets.

[1] Web of Data - (http://richard.cyganiak.de/2007/10/lod/)

Our study is therefore closer to the works that are focused on SPARQL query rewriting and reformulation. Although we present significant innovations in a domain such as the semantic web, in which has hardly developed studies on this topic. Makris et. al.[5] is quite close to our approach. A formal model for RDF triple patterns rewriting is defined. A quite expressive specific mapping language based on Description Logics constructs is defined and used for the query rewriting. Nevertheless, the rewriting of triple patterns is not dependant on mapping relationships (i.e. equivalence or subsumption). These relationships affect only the evaluation results of the rewriten query over the target ontology. Therefore, they do not take into account the estimation of loss in precision or loss in recall. Moreover, it is not clear what is done when there are not enough mapping expressions to rewrite every term of the source query.

On query relaxation field, there are studies like Hurtado et al. [3], where a new clause of SPARQL, called "RELAX", is introduced for make queries more flexibles by a logical relaxation of the conditions enclosed by the clause. This approach is far from our study, because they are not focused on translating the entire query and extend it with other data sources, but in generalize some conditions of it into the same dataset.

Outside the areas of query rewriting or relaxation, Herzig's article [2] presents similar objectives to ours, regarding the goal of query reusing for consult additional datasets. They make a ERM(Entity relevance model) that contains the structure and content of the results needed to answer a query and thus it can be used to transfer the query to other datasets. It has the disadvantage of allowing only the query of entities.

Finally a differentiating aspect of our system is the measure of the loss of information. For compute it we adapt the approach presented by Salton [7] to estimate the information loss when a term is substituted by an expression. We use the metrics precision and recall originating from Information retrieval [9], [1]. There are other metrics like similarity [4], distance between two ontology concepts, that we are studying to adapt too to our approach.

3 Proposed approach

My purpose is to exploit RDF-ied sources, being they native RDF Linked Open Data sources or having an RDF scheme wrapping with non RDF data source (e.g. relational database with appropriate RDF scheme mapping). Once the original SPARQL query is received, a SPARQL query engine is launched on the by default dataset. After receiving the answer, there is the possibility to ask for more answers. In that case the original query can be sent to different data sources that share the vocabulary used in the query. But if that chance is not available or its answers are not enough, then a query rewriting process begins. Different choices are possible depending on the user decision: 1) to rewrite the original query (slightly changing its semantics) over the same source but looking for different answers to those previously obtained, 2) to try to rewrite the query using another related source with different vocabularies according to the knowledge

managed by our system, and 3) to ask to the user to select another source from a list offered by our system.

All of the choices take advantage of semantic relationships, already existing and accessible by our system, associated to the terms appearing in the original query. Term semantic relationships include but are not necessarily limited to synonymy, hyponymy, and hyperonymy (for instance, consider meronymy). Those relationships may be taken from repositories such as VoID[2] linksets or tools like WordNet[3] or can be described into the RDF datasets. Changing a term for a related one, derives in a change of semantics. The challenge is to be able to appropriately measure that change in order to assist the user when informing with the answers.

In a first attempt query rewriting can be approached term by term. Then, when all the terms of the original query are rewritten we say we have a complete translation (notice that it may also incurred in a semantic change). When there are terms in the original query without associated semantic relationships, we say we have a partial translation. A significant challenge is to manage how to cope with such a scenario. Different approaches are possible. For instance, try to find a translation for the union of its registered hyponyms, or try with the conjunction of its registered hyperonyms. In any case, measures for precision and recall for the query translation must be developed. Using such metrics a user is allowed to establish a threshold for the admitted loss in precision or loss in recall estimated for the received answers. For example, if the user defines a limit of 20% the system must guarantee that the amount of unwanted (loss in precision) or missed data (loss in recall) in the future answers presented to the user is kept always below 20% of the information showed. Moreover, the rewriting approach can be enhanced by allowing the rewriting of query expressions (instead of only single terms).

4 Methodology and schedule

Our research can be scheduled into three phases.

In the *first phase* I have analyzed related works in the field of SPARQL query engines as well as works that consider query approaches on the database area, taking into account the query rewriting and relaxation techniques. Once I identified their contributions and weaknesses I defined a global architecture of my proposal with an specification of the functionalities of the modules that constitute that architecture. In the *second phase* I am concentrating my efforts on providing an innovative solution for the following two aspects:

- **Rewriting process**. I am developing an algorithm that tries to rewrite the query in order to get a complete translation of it, and if that is not possible in order to get a partial translation. Different strategies are possible to search for translations.

[2] VoID - (http://www.w3.org/TR/void/)
[3] WordNet - (http://wordnet.princeton.edu/)

– **Definition of metrics** that allows to estimate the information loss when semantic equivalence of the original query is not preserved.On this stage, we review the different metrics from Information Retrieval and their literature. Later we adapt the selected metrics to our approach.

In the *third phase* implementations for all those processes will be deployed and proper experimentation will be performed to test the approach.

5 Conclusion

In this paper we propose an approach for the federated query processing of heterogeneous Linked Data sources, based on the rewriting of the initial user query into new queries formulated in terms of the target data sources. To perform this task we are developing a new translation algorithm that uses ontology mapping and query rewriting techniques. Our final aim is to enrich the answer in an incremental manner with data obtained by querying each time to a different datasource, measuring the possible loss of information if semantic changes are detected in the reformulated query.

6 Acknowledgements

This work is supported by the TIN2010-21387-CO2-01 project.

References

1. Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 348–353, 2007.
2. D Herzig and Thanh Tran. One query to bind them all. In *COLD2011, CEUR Workshop Proceedings*, volume 782, 2011.
3. Carlos Hurtado, Alexandra Poulovassilis, and Peter Wood. Query relaxation in rdf. *Journal on data semantics X*, pages 31–61, 2008.
4. A Maedche and S Staab. Measuring similarity between ontologies. *Knowledge engineering and knowledge management: Ontologies and the semantic web*, pages 15–21, 2002.
5. Konstantinos Makris, Nektarios Gioldasis, Nikos Bikakis, and Stavros Christodoulakis. Ontology mapping and sparql rewriting for querying federated rdf data sources. *On the Move to Meaningful Internet Systems, OTM 2010*, pages 1108–1117, 2010.
6. Bastian Quilitz and Ulf Leser. Querying distributed rdf data sources with sparql. *The Semantic Web: Research and Applications*, pages 524–538, 2008.
7. Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of*. Addison-Wesley, 1989.
8. Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt. Fedx: Optimization techniques for federated query processing on linked data. *The Semantic Web–ISWC 2011*, pages 601–616, 2011.
9. Cornelis Joost Van Rijsbergen. *Information Retrieval. Evaluation - (http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html)*.