SemaGrow: Data Intensive Techniques to Boost the Real-Time Performance of Global Agricultural Data Infrastructures

Angelos Charalambidis and Stasinos Konstantopoulos

Institute of Informatics and Telecommunications, NCSR "Demokritos",
Aghia Paraskevi 153 10, Athens, Greece
{acharal,konstant}@iit.demokritos.gr

1 SemaGrow Objectives and Tools

As the trend to open up data and provide them freely on the Internet intensifies, the opportunities to create added value by combining and cross-indexing *heterogeneous* data at a *large scale* increase. To seize these opportunities, we need infrastructure that is not only efficient, real-time responsive and scalable, but is also flexible and robust enough to welcome data in any schema and form and to transparently relegate and translate queries from a unifying end-point to the multitude of data services that make up the open data cloud. FP7-ICT SemaGrow develops novel technologies for creating this infrastructure, including *dynamic data integration* and *distributed querying* technologies specifically designed for today's cloud of large, heterogeneous, live and constantly updated datasets.

What is important to note is that the SemaGrow architecture emphasises querying efficiency without requiring cloning or any modification of the federated endpoints and existing workflows.

The focal point of this infrastructure is the *SemaGrow Stack*, providing to data consumers a SPARQL endpoint that federates multiple SPARQL endpoints independently maintained by data providers. The SemaGrow Stack provides a querying interface that uses dynamic vocabulary transformation in order to apply the results of *ontology alignment* and address the fragmentation of the vocabularies used in the federated data sources.

It also applies methods from distributed databases in order to construct *efficient distributed querying strategies*, dynamically adapting to currently observed latency and throughput, and foreseeing mechanism for falling back to alternatives in the face of unavailability. These methods take into account complex and inter-dependent considerations involving data contents and querying efficiency. These considerations stem from manually provided descriptions of the federated data sources, previously noted latency and throughput observations, and statistics extracted from observing the flow of query responses from the federated endpoints back to the client application.

2 Intelligent Query Decomposition Prototype

The focus of this demonstration is the intelligence behind transparent data integration: the SemaGrow Stack constructs *query strategies* that detail which patterns of the overall query should be executed at each of the endpoints known to the federation, including alternatives in the face of endpoint unavailability. These strategies are optimized based on *data source metadata* regarding the *reactivity and availability track record* of each endpoint, its *contents* and the *schemas* they follow, and known *alignments* between these schemas. This metadata is partially authored and partially computed, including descriptions provided by human data curators, statistics extracted by analysing measurements obtained during query execution, and knowledge extracted by *ontology alignment*.

The optimal strategy, including how to break up the query into subqueries and which federated endpoint should each subquery be executed at, is computed automatically by the query optimizer. The optimizer searches among the possible query decompositions and chooses an optimal execution plan with respect to the overall cost given by a cost function. The cost function takes into account the execution and the communication cost of each subquery and the transformation and merging cost of the individual query answers.

During the demonstration we use a graphical tool to update source descriptions and discuss the query strategy changes affected by these updates.

The schema description contains both structural and quantitative information about the data source. For example, the data curator can inform about the entities that are described in the data source by providing regular expressions that succinctly delineate the URI space within which these entities reside.

3 Expected Synergies

SemaGrow outcomes are validated in use cases from the agro-environmental research community, where *data-intensive analysis and modelling* needs to to combine information from many, usually large and heterogeneous, actively maintained sources.

Our objective is to explore the potential for cooperation with projects and activities that are active in *benchmarking* distributed data management solutions, especially over loose federations where data heterogeneity and end-point unreliability are taken into consideration.

A further objective is to explore opportunities for pushing our notion of data heterogeneity beyond schema heterogeneity to also include data and knowledge from numerical and geo-spatial databases. SemaGrow technologies for intelligent query strategy construction can not only be transferred to different data management paradigms, but also generalized into a unifying framework for aggregating big data solutions.

Acknowledgements The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 318497. Please also see `http://www.semagrow.eu`