Enabling Semantic Integration of Streaming Data Sources

Jean-Paul Calbimonte

Ontology Engineering Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid,
Campus de Montegancedo s/n 28660, Boadilla del Monte, Spain
jp.calbimonte@upm.es

Abstract. We propose a distributed ontology-based approach for integration of streaming live data sources, using extensions of SPARQL for streams and declarative mappings for query rewriting. The challenge of exposing live data from streams such as those provided by sensor networks, using semantically rich models and queries is becoming more relevant nowadays. The goal of this PhD thesis[1] is to investigate, analyse and propose solutions that bridge the gap between semantic data access, streaming query evaluation and data integration.

1 Problem Statement

Nowadays in scientific and industrial environments, the amount of data in form of heterogeneous streams is becoming one of the main sources of information and knowledge acquisition. Advances in wireless communications and sensor technologies have enabled the deployment of networks of interconnected devices capable of ubiquitous data capture, processing and delivery of such streams.

Many solutions for accessing and processing this type of data have been devised in the last decade[3, 1, 9, 15, 11, 8, 2], including acquisitional and non-acquisitional streams. However the management and querying techniques explored so far have not provided consistent solutions for the problem of semantic heterogeneity of streaming data sources. This problem has become even more evident given the increasing number of sensor infrastructures in several domains, each one having completely different schemas, stream rates, quality of service and delivery mechanisms.

On the other hand the approaches for semantic data provision have largely focused on stored data[16]. And only very few solutions have been proposed for querying streams using semantic technologies [5, 7]. A contribution on this area is becoming an imperative need in the scope of the recent efforts of making information accessible to the web of data, and in particular from the increasing number of streaming sources such as sensor networks. Having such a information space would enable applications to obtain live data published by third-party providers regardless of formats, schemas, rates and underlying implementation.

[1] The author completed the first year of the thesis (Phase I), under the supervision of Oscar Corcho at Universidad Politcnica de Madrid.

2 Main Research Questions

Our work is centred on the question of how to integrate both streaming and stored information sources using rewriting techniques and how to expose a consistent view of the continuously generated data as RDF for the semantic web. This will also prompt us to identify the suitable language extensions to existing approaches that take into account streaming operators in SPARQL queries [5, 7]. In this context we are also interested in finding appropriate means to represent the mappings and correspondences between ontological elements and stream/relational elements, in a way that can be reused for query rewriting from SPARQL queries to native streaming queries. In this scope we also ask how we can optimise the query rewriting approach so that it is possible to efficiently query live streaming data for the original sources, without increasing the time response beyond acceptable time frames.

By tackling these issues we intend to provide the foundations of an infrastructure that is able to provide live data though SPARQL-extended queries with stream capabilities, not from one but several and heterogeneous sources that are mapped to ontological views. An integrated and distributed platform of such characteristics has not been devised yet and we believe that it can have a certain impact on the community. Nevertheless there is a number of problems that we will leave for future or complementary research. Namely the ability to automatically generate mappings or the discovery of a-priori unknown streaming data sources.

3 General Approach

Our work is organised in the following phases that cover the planned research.

Related work First we have studied the literature and existing approaches for the main topics that we cover in this thesis: relational-to-ontology data access, querying RDF streams, stream management systems and distributed query processing. Understanding the base approaches will help us coming up with a solution to the problems we are interested in, and knowing what techniques and technologies we can reuse or base upon for our approach.

Ontology-based data access Once we have studied the previous works we have presented a first approach for ontology-based data access to streams [10]. This consists of a simple mechanism of using mapping assertions based on the $\mathsf{R_2O}$[6] language that relate stream elements to ontology elements. Then an extended version of SPARQL with extensions for streaming queries (based on the C-SPARQL language[5]) is used to issue queries over the ontological view and a translation component transforms them to stream queries in a language that can be executed by a stream management system (such as SNEE[12], STREAM[3], etc.). See Section 4 for details.

Stream integration After the data access step we plan to propose an approach for the integration of heterogeneous streaming data sources. This includes the mapping relationships from both stream and stored sources to ontological views,

and the rewriting of queries to a distributed query processor that is able to execute and integrate data from the streaming sources. We need to formally specify the algebra to which our streaming queries will be transformed, and indicate the optimisation algorithms that we will employ or design, including classical pushing of operators, optimisations of joins and query planning tasks. A service implementing this integration & query functionality will be provided at the end of this stage.

Evaluation & optimisation Once we have the core component of the thesis, we will evaluate the approach from three main perspectives: the query execution response time and overhead of the rewriting process; the expressiveness of queries rewritten from extended-SPARQL to native streaming languages; the expressive power of mappings from streaming elements to ontology elements. We plan using known benchmarks for streaming data sources[4] (See Section 5). We will iteratively use optimisation techniques and progressively evaluate the results that we obtain during this phase. The resulting prototype will provide a distributed integration service for heterogeneous streaming and stored data sources using SPARQL with streaming extensions (named $\text{SPARQL}_{\text{STR}}$ in this paper) as a query language, providing live data for the semantic web.

4 Proposed Solution

Our approach consists in creating an Ontology-based streaming data integration service (Fig 1) that can receive requests over a global ontological view in $\text{SPARQL}_{\text{STR}}$. This global ontology can optionally be aligned and represented in terms of other ontologies that the sources are mapped to. This can be done using Ontology-to-Ontology mappings that can be partially computed using alignment technologies. Using these correspondences the original query can be rewritten (query reconciliation) if necessary. Sources can be exposed in terms of ontological views using a set of Ontology-to-Source mappings, which correlate ontology entities and stream/stored entities. These are provided with a-priori knowledge of the ontologies and sources schemas, and are based on the $\mathsf{R_2O}$[6] mapping language, which has been extended to support streaming queries and data.

So when a $\text{SPARQL}_{\text{STR}}$ query arrives, it is translated (query translation) to an internal algebra that is capable of dealing with streaming and stored sources. This transformation is made using the already mentioned Ontology-to-Source mappings. The algebra is an extension of the one defined for SNEEql[9], a continuous query language that has expressive window and window-to-stream operations, and a semantics that incorporates both streaming and stored data.

Once the transformations are made, the distributed query processor for streams is in charge of the logical rewriting and physical optimisations that take into account rewriting rules (push of operators, order of joins, etc) and cost models to find the best distributed query plan that will be executed by the evaluator, dispatched to the participating processors and then integrated[14]. Note that the execution in sources such as sensor networks may include in-network query processing, pull or push based data delivery and other data source specific settings.

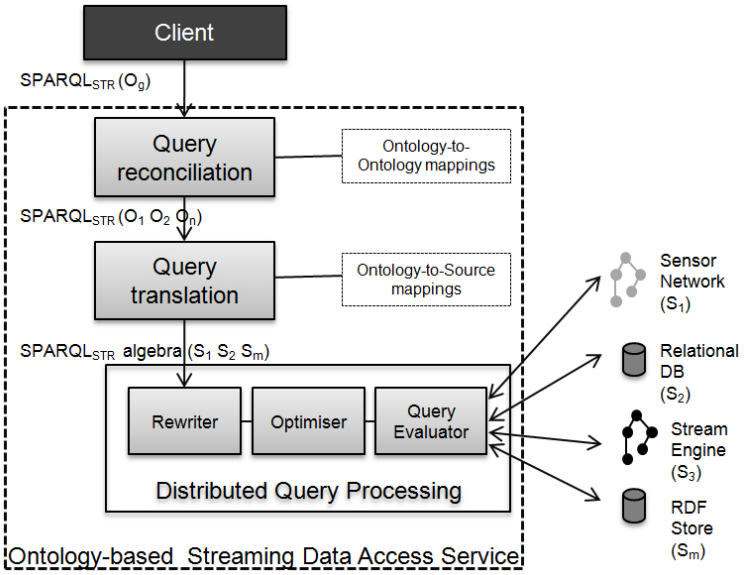

Fig. 1. Ontology-based Streaming Data Access service

5 Evaluation

In order to evaluate the proposed solution we will first identify the main targets of the evaluation, which will be focused in the following points: performance of the query rewriting, performance of the query execution, expressive power of the mappings, expressiveness of the query language. For this purpose we will use the well-known Linear Road benchmark to compare our results to non-ontology-based approaches in terms of performance and expressiveness. Moreover and in order to validate the usability of our approach to real applications we will use our proposal in order to provide integrated streaming and stored data sources for the use cases of the SemSorGrid4Env[13] project for environmental monitoring. These include a Coastal and Estuarine Flood Warning system in southern UK and a Fire Risk Monitoring and Warning system in northern Spain.

6 Future Work

Although we have shown initial results querying the underlying streaming engine with basic queries[10], we expect to consider in the near future more complex query expressions including aggregates, and joins involving both streaming and stored data sources. Another important strand of future work is the optimisation of distributed query processing [14] and the streaming queries [1, 12]. In the scope of a larger streaming and sensor networks integration framework, we intend to achieve the following goals: i) integrating streaming and stored data sources through an ontological unified view; ii) combining data from event-based and acquisition-based streams, and stored data sources. The present work can be considered as a first step to our goal of providing an ontology-based integration platform for continuous heterogeneous data sources.

Acknowledgments. This work is supported by the European Commission project SemSorGrid4Env (FP7-223913).

References

1. Abadi, D.J., Ahmad, Y., Balazinska, M., Cetintemel, U., Cherniack, M., Hwang, J.H., Lindner, W., Maskey, A.S., Rasin, A., Ryvkina, E., Tatbul, N., Xing, Y., Zdonik, S.: The Design of the Borealis Stream Processing Engine. In: CIDR (2005)
2. Aberer, K., Hauswirth, M., Salehi, A.: A middleware for fast and flexible sensor network deployment. In: VLDB. pp. 1199–1202. ACM (2006)
3. Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Ito, K., Motwani, R., Srivastava, U., Widom, J.: Stream: The stanford data stream management system. Tech. rep., Stanford InfoLab (2004)
4. Arasu, A., Cherniack, M., Galvez, E., Maier, D., Maskey, A.S., Ryvkina, E., Stonebraker, M., Tibbetts, R.: Linear road: a stream data management benchmark. In: VLDB 2004. pp. 480–491. VLDB Endowment (2004)
5. Barbieri, D.F., Braga, D., Ceri, S., Grossniklaus, M.: An execution environment for C-SPARQL queries. In: EDBT 2010. pp. 441–452 (March 2010)
6. Barrasa, J., Corcho, O., Gómez-Pérez, A.: R2O, an extensible and semantically based database-to-ontology mapping language. In: SWDB2004. pp. 1069–1070 (2004)
7. Bolles, A., Grawunder, M., Jacobi, J.: Streaming SPARQL - extending SPARQL to process data streams. In: ESWC 08. pp. 448–462 (2008)
8. Botan, I., Cho, Y., Derakhshan, R., Dindar, N., Gupta, A., Haas, L.M., Kim, K., Lee, C., Mundada, G., Shan, M.C., Tatbul, N., Yan, Y., Yun, B., Zhang, J.: A demonstration of the maxstream federated stream processing system. In: ICDE. pp. 1093–1096. IEEE (2010)
9. Brenninkmeijer, C.Y., Galpin, I., Fernandes, A.A., Paton, N.W.: A semantics for a query language over sensors, streams and relations. In: BNCOD '08. pp. 87–99 (2008)
10. Calbimonte, J.P., Corcho, O., Gray, A.J.G.: Ontology-based Access to Streaming Data. In: Poster paper at ESWC 2010 (2010)
11. Chandrasekaran, S., Cooper, O., Deshpande, A., Franklin, M.J., Hellerstein, J.M., Hong, W., Krishnamurthy, S., Madden, S.R., Reiss, F., Shah, M.A.: TelegraphCQ: continuous dataflow processing. In: SIGMOD '03. pp. 668–668 (2003)
12. Galpin, I., Brenninkmeijer, C.Y., Jabeen, F., Fernandes, A.A., Paton, N.W.: Comprehensive optimization of declarative sensor network queries. In: SSDBM 2009. pp. 339–360 (2009)
13. Gray, A.J.G., Galpin, I., Fernandes, A.A.A., Paton, N.W., Page, K., Sadler, J., Koubarakis, M., Kyzirakos, K., Calbimonte, J.P., Corcho, O., García, R., Díaz, V.M., Liebana, I.: SemSorGrid4Env architecture phase i. deliverable. d1.3v1 (2009)
14. Kossmann, D.: The state of the art in distributed query processing. ACM Comput. Surv. 32(4), 422–469 (2000)
15. Madden, S.R., Franklin, M.J., Hellerstein, J.M., Hong, W.: TinyDB: an acquisitional query processing system for sensor networks. ACM Trans. Database Syst. 30(1), 122–173 (2005)
16. Sahoo, S.S., Halb, W., Hellmann, S., Idehen, K., Jr, T.T., Auer, S., Sequeda, J., Ezzat, A.: A survey of current approaches for mapping of relational databases to RDF. W3C (January 2009), `http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf`