Service-Oriented Development of Workflow-Based Semantic Reasoning Applications

Alexey Cheptsov, High Performance Computing Center Stuttgart (HLRS), Stuttgart, Germany

Stefan Wesner, Institute of Organisation and Management, University of Ulm, Ulm, Germany

Bastian Koller, High Performance Computing Center Stuttgart (HLRS), Stuttgart, Germany

ABSTRACT

The modern Semantic Web scenarios require reasoning algorithms to be flexible, modular, and highly-configurable. A solid approach, followed in the design of the most currently existing reasoners, is not sufficient when dealing with today's challenges of data analysis across multiple sources of heterogeneous data or when the data amount grows to the "Big Data" sizes. The "reasoning as a workflow" concept has attracted a lot of attention in the design of new-generation Semantic Web applications, offering a lot of opportunities to improve both flexibility and scalability of the reasoning process. Considering a single workflow component as a service offers a lot of opportunities for a reasoning algorithm to target a much wider range of potentially enabled Semantic Web use cases by taking benefits of a service-oriented and component-based implementation. We introduce a technique for developing service-oriented Semantic Reasoning applications based on the workflow concept. We also present the Large Knowledge Collider - a software platform for developing workflow-based Semantic Web applications, taking advantages of on-demand high performance computing and cloud infrastructures.

Keywords: High Performance Computing, Parallelization, Reasoning, Semantic Web, Service, Workflow

INTRODUCTION

The large- and internet-scale data applications are the primary challenger for the Semantic Web, and in particular for reasoning algorithms, used for processing exploding volumes of data, exposed currently on the Web. Reasoning is the process of making implicit logical inferences from the explicit set of facts or statements, which constitute the core of any knowledge base. The key problem for most of the modern reasoning engines, e.g., Jena, Pellet, etc., is that they cannot efficiently be applied for the real-life data sets that consist of tens, some-

DOI: 10.4018/ijdst.2014010103

times of hundreds of billions of triples (a unit of the semantically annotated information), which can correspond to several Petabytes of digital information stored on a disc. Whereas modern advances in the supercomputing allow this limitation to be overcome, the reasoning algorithms' logic need to be adapted to the demands of rapidly growing data universe, in order to be able to take advantages of large-scale and on-demand infrastructures such as high performance computing or cloud technology. On the other hand, the algorithmic principals of the reasoning engines need to be reconsidered as well in order to allow for very large volumes of data. Service-oriented architectures (SOA) can greatly contribute to this goal, acting as the main enabler of the trend-new reasoning techniques, such as incomplete reasoning, introduced by (Fensel & Van Harmelen, 2008). This paper discusses an approach for service-oriented development of workflow-driven semantic reasoners, focusing on performance-critical application scenarios. The approach is based on the Large Knowledge Collider (LarKC) paltform, introduced in our previous publication (Cheptsov & Wesner, 2012).

The paper is organized as follows. In the first section – "Towards semantic reasoning on the web scale", we collect our consideration towards enabling the large-scale web reasoning for Semantic Web applications. In the second section – "Large Knowledge Collider - making the semantic reasoning more service-oriented", we discuss LarKC – a service-oriented platform for development of fundamentally new reasoning application, with much higher scalability barriers as by the existing solutions. In the third section – "Success stories and application examples", we introduce some successful applications implemented with LarKC, such as Bottari – the Semantic Challenge winner in 2011. In the last section – "Conclusion and future work", we summarize main results presented in the paper and highlight the directions of future work in highly-scalable semantic web reasoning.

TOWARDS SEMANTIC REASONING ON THE WEB SCALE

From Web to the Semantic Web

The Web as it is seen by the users "behind the browser" has traditionally been one of the most successful examples of the SOA realization. The possibility to transform the application's business logic into a set of the linked services supplied with the transparent access to those services over standardized protocols such as HTTP was a key asset for tremendous widespread of the Internet worldwide. However the possibility to organize business relationship between the data located on several hosts had been extremely poor. The research seeking for a concept of applying a data model on the Web scale resulted in the Semantic Web – the later advance of the Web, which offers a possibility to extend the Web-enabled data with the annotation of their semantics, thus making the context in which the data is used meaningful for the applications, as elaborated by (Broekstra et al., 2001). Nowadays, there are several existing well-established standards for annotation of data web-wide, such as for example Resource Description Framework (RDF) schema.

The practical value of the Semantic Web is to enable the development of applications that can handle complex human queries based not only on the value of the analyzed data, but also on its meaning. Promotion of such platforms as (Friends-of-a-Friend) FOAF[1] at the early stages of the Semantic Web has forced a lot of data providers to actively expose and interlink their data on the Web, which resulted in many problem-oriented data repositories, as for example Linked Life Data (LLD)[2], which is a collection of the data for biomedical domain; alone the LLD dataset comprises over one billion web resources presented in RDF. On the other hand, social networks like Twitter or Facebook encourage people to upload there personal data