Semantic Web and Retrieval of Scientific Data Semantics

Goran Soldar
School of Computing and Mathematical Sciences
University of Brighton
g.soldar@brighton.ac.uk,

Dan Smith
School of Information Systems
University of East Anglia
djs@sys.uea.ac.uk

Project Description

Introduction

The results of scientific activities that include observations, experiments, interactions, deductions, etc. are stored in data sets and often made available to the scientific community. These large volumes of scientific data are typically kept and managed in an ad-hoc manner and it requires a substantial effort to discover and evaluate data quality and suitability for a particular analysis. The provision of appropriate metadata and links to related resources enables the possibility of intelligent assistance in finding and evaluating data sets for scientific research. Such tools are of increasing importance as the volume of scientific data is increasing rapidly due to the extensive deployment of automated data collection and monitoring systems. To find and process scientific data sets available on the Web is time consuming despite the fast and powerful search engines. For example to find files related to temperatures we used the Google search engine. The key word "Temperature" was entered and search restricted to the University of East Anglia Climate Research Unit (`http://www.cru.uea.ac.uk`). Google returned 773 pages. For a scientist browsing every single web page to find required data sets, without guarantee that they will find what they need, is simply too much time consuming exercise. It would be desirable that data files are described in a way that they could be quickly found and their semantics learnt automatically by a machine. The framework for such a semantic retrieval of scientific data is offered by W3C in the form of Resource Description Framework (RDF) [1,2].

This project seeks to address methods and provide an architecture for describing, managing, retrieving, and extracting semantic information from specific science domains. The conceptualization of the subject domains including the development of the vocabulary [3], is performed using RDF model and schema constructs. We distinguish between two levels of metadata related to RDF, *Instance Metadata* and *Schema Metadata*. The description of a particular resource using RDF constructs is known as the instance metadata, whereas Schema Metadata describes a particular ontology The notion of a semantic case is being introduced to capture semantic forms. The heterogeneity of data sources is exposed and integration of data is achieved through a mediation process [4]. The primary goal of our project is to enable machine processing of semantic information that would be utilized to query the actual information from data sets To address the problem of extracting semantic information from data files, we build an ontology for the meteorology domain which is then use to create semantic cases as file description templates. We use RDF Model Syntax and RDF Schema to create semantic cases instances. The architectural infrastructure is based on the Semantic Retrieval Model [5].

Storage and Management of Ontologies

The management of RDF structures is an open issue, and W3C does not recommend any particular method for manipulating such data. By definition RDF Model is set of triples. This property can be utilised to achieve manipulation of RDF triples as a Relational Model. The storage and manipulation of RDF graph structures represents a problem. It does not matter whether the data is kept in RDF syntax or as triples, the problem is that there is no RDF data management system that provides access and manipulation of ontologies. In addition the existing RDF parsers are design to work on single documents only, so the problem is how to access documents remotely, and how to achieve the maintenance either of parts or of the whole documents in terms of modification, insertion, and deletion. Another question that arises is whether there is a need to keep whole RDF documents at all. Since modern database management systems are based on the concept of the Relational Model it is considered to use the existing RDBMS for manipulation of RDF data or alternatively to build RDF native data storage model.

Retrieval and Extraction of Semantics (RDF Triple Engine)

The system is based on client-server architecture that includes specialized RDF servers. The advantage of this approach is in data management facilities of RDBMS that are utilized for the manipulation of raw RDF data. The architecture comprises the 4 layers: 1) *Interface Layer*, 2) *Web Infrastructure Layer*, 3) *RDF Management Layer* and 4) *DBMS Layer* The interface layer provides access to semantics descriptions to human users as well as to application programs. The prototype of the Semantics Retrieval Language (SRL) is currently being developed to provide semantics-orientated retrieval of DBMS managed RDF data. QBE-like Web form is used to assist users in specifying their requests although SQL-like statements are also supported. The interface handlers (access rights and authentication) are built as Java servlet modules and they run inside the Apache Tomcat servlet engine, which is seamlessly connected to the Apache Web server. All valid requests are transparently forwarded to the Semantics Support Server (SSS). Since SSS runs as a separate service, applications can establish direct connection with this server using the internal SRL protocol over TCP/IP connection.

RDF Triples Engine (RTE) is a module responsible for manipulating triples and executing semantic queries. The check for namespace existence for each prefix is performed by RTE before the records are inserted into the database. A user with no knowledge of graph structures and its specific elements (containers) will find it difficult to assemble the full picture from the output. This problem is solved by adding the additional query processing to RTE, which is aware of RDF semantic and also is able to produce results suitable for human use (HTML document that conforms to XHTML structure) or for further processing (XML structure).

References

[1] WWW Consortium "Resource Description Framework (RDF), Model and Syntax Specification", Available at: *http://w3.org./TR/REC-rdf-syntax.*
[2] WWW Consortium "Resource Description Framework (RDF) Schema Specification", Candidate Recommendation, Available at: *http://w3.org./TR/1999/REC-rdf-schema.*
[3] T. Gruber “Towards Principles for the Design of Ontologies Used for Knowledge Sharing”, *International Workshop on Formal Ontology, Padova, Italy*, 1993.
[4] D. Smith, M. Lopez, “Information finding and filtering for collections of semi-structured documents”, *Proc. INFORSID XV*, Toulouse, 353-367, 1997.
[5] G. Soldar, D. Smith “Retrieving Semantics from Scientific Data: A Domain Specific Approach Using RDF”, *Proceedings of the IASTED International Symposia on Applied Informatics*, Innsbruck, Austria, February 2001