OSDC/PIRE research proposal

Service-ware Research Group, Information Technology Research Institute, AIST

The Service-ware Research Group focuses on research into the design, deployment and integration of cloud-based services with an emphasis on data management and data integration. Examples of current projects include large scale analytical processing of Big Data (including satellite data archives from GEO Grid [http://www.geogrid.org/]) and database processing over Linked Data [http://linkeddata.org].

Analysis of Linked Data to support live query processing

Introduction

Although indexes exist such as Sindice [http://sindice.com] support efficient query answering over cached parts of the Semantic Web, emerging live decentralised approaches can provide fresher results by accessing Linked Data directly at query execution time. At AIST we have developed an adaptive query processor for RDF, ADERIS [http://code.google.com/p/sparql-aderis/], to which we have recently added support for such “live” query processing. The proposed project involves performing a detailed analysis of Linked Data on the Web to develop a predictive model to rank the priority with which Linked Data should be investigated at runtime by ADERIS.

Research Goals

1) Select, retrieve and analyse Linked RDF data (e.g. a subset of the Billions Triples Challenge data set [http://km.aifb.kit.edu/projects/btc-2012/]) to develop a predictive model to be used by ADERIS when deciding on whether to attempt to dereference a link to RDF data encountered during the execution of a query. The model could be based on a Naive Bayesian Classifier or something more complex. Processing a large enough sample will require parallel execution, for example using Hadoop, during which RDF is extracted from data sources and analysed.

2) Produce a (preferably interactive) summary of the data, e.g. using the Google Visualisation API or other tools, with the goal of presenting interesting statistics and trends to Semantic Web researchers.

Knowledge base and/or skills to be achieved

An interest in Semantic Web and large-scale data mining/analysis and the ability to solve practical issues when designing and running parallel data mining tasks (e.g. using a Hadoop cluster) to access large amounts of RDF data on the Web.