The Social Intelligence Benchmark

Minh-Duc Pham
CWI, Amsterdam
<duc@cwi.nl>

Peter Boncz
CWI, Amsterdam
<boncz@cwi.nl>

Abstract

The Resource Description Framework (RDF) and its recommendation query language, SPARQL, have recently been gaining traction as the main semantic web technology for publishing and integrating large collections of interlinked datasets over the web. Subsequently, a range of RDF stores has been released such as Virtuoso, OWLIM, Jena TDB, Bigdata and Stardog. In order to evaluate the performance of such systems, a number of RDF/SPARQL benchmarks have been proposed. However, these existing benchmarks are often limited in representing real datasets or mostly relational-like. While real-life RDF data is irregular in shape, highly correlated with skewed data distributions, these existing benchmarks contain no such correlations or irregularities in both their structure and data values, which are generated with simple uniform data distributions. Moreover, as the data models and the logical schemas of these benchmarks can be easily represented in the relational model as well, they hardly highlight the advantages of using RDF/SPARQL.

In this work, we propose a graph database benchmark called the Social Intelligence Benchmark (SIB) that models a large social network, whose data is combined with Linked Open Data (LOD) sources. Popular web-scale social networks such as Facebook and Twitter can be seen as managing a huge highly-connected data graph, consisting of entities like users, photos, annotations and posts that are interlinked using friendships, tagging and discussions. In recent years, a flurry of new start-up companies providing systems to manage large graph datasets have appeared, such as Neo4j, InfinityGraph, GraphDB, AllegroGraph and Dydra. The latter two systems are based on RDF, which is in essence a graph data model. With the SIB we thus aim to test database systems, focusing on RDF, in their capability of managing a potentially huge graph data structure, with correlated structure and values, modeled after a social network.

The main goal of the benchmark is to measure database system performance and scalability in an area where the graph data model of RDF could provide advantage over the relational model, with four targets: (i) to test graph database performance on a highly connected graph where materialization of a closure is impossible, (ii) to do so in the recognizable usage scenario of a social network, both mimicking the day-to-day lookup and update queries generated by users to perform actions in the social network as well as complex analytical queries over their behavior, (iii) to make data and queries more realistic by avoiding uniform distributions in the generated data and include correlations between graph connectivity and user characteristics and (iv) to add real (non-synthetic) RDF data to the benchmark by linking the social graph to well-known linked open datasets, in particular DBpedia and LinkedGeodata.

Efficiently generating a huge highly-connected graph, much larger than RAM, whose shape and data values are semantically correlated, is non-trivial. Therefore, apart from outlining the generated dataset and queries that make up the benchmark, we also describe in detail the scalable algorithm we devised to generate this complex dataset on a cluster of computers.