Karlsruher Institut für Technologie

Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Graduiertenkolloquium Angewandte Informatik

„Query Processing over Graph-based Data on the Web"

M.Sc. Maribel Acosta

AIFB

Abstract:
Linked Data initiatives have encouraged the publication of large datasets in the Linking Open Data (LOD) cloud, where data is represented using the graph-based RDF data model. In order to support querying capabilities over Linked Data sets, Web access interfaces such as SPARQL endpoints or Triple Pattern Fragment (TPF) servers have been deployed. TPF servers support the evaluation of single triple patterns and have been recently proposed as a highly available mechanism to query RDF data online. Despite these developments, the Web-like characteristics of Linked Data sources pose fundamental challenges to Linked Data management, specifically on remote query processing over RDF datasets. The lack of statistics about selectivities and data distributions, as well as unpredictable data transfer rates and server workload, can negatively impact the effectiveness of query engines against Linked Data, even in presence of the innovative querying capabilities offered by TPF servers. This problem is mainly generated because existing SPARQL query engines implement query execution strategies that rely in some way on the traditional optimize-then-execute paradigm, instead of following adaptive strategies that adjust query executions to unexpected data source conditions. In this talk, we present an adaptive SPARQL query engine tailored to TPFs servers that relies on a network of Linked Data Eddies (nLDE). Our proposed solution is able to adjust query execution schedulers to data availability and runtime conditions. We will present the results of our experimental studies which suggest that nLDE outperforms static Web query schedulers in scenarios with unpredictable transfer delays and data distributions.

An orthogonal, but equally important aspect of querying Linked Data is the quality of the retrieved data. Recent studies reveal that RDF datasets exhibit varying quality in different dimensions including completeness, semantic validity, and semantic accuracy. Moreover, the semi-structured nature of RDF data makes it very hard to assess the quality of datasets up front. Executing SPARQL queries against data with quality issues leads to low-quality and even incomplete results. To overcome this limitation, we present HARE, a hybrid query processing engine that brings together machine and human computation to execute SPARQL queries. HARE extends the optimization techniques presented in nLDE to identify efficient hybrid query plans. Furthermore, HARE implements a novel query engine that relies on the graph topology of RDF data to decide on-the-fly which parts of a SPARQL query should be executed against a dataset or via crowdsourcing. In addition, HARE exploits knowledge encoded in the RDF graphs to create questionnaires that can be answered by the crowd. In this talk, we will discuss the results of an empirical evaluation of HARE, and show how HARE is able to accurately enhance SPARQL query answer completeness.

Termin: **Mittwoch, 08. Juni 2016, 15.45 Uhr**
Ort: **Englerstraße 11, 76131 Karlsruhe**
Kollegiengebäude am Ehrenhof (Geb. 11.40), 2. OG, Raum 231
(Hinweise für Besucher: www.aifb.kit.edu/web/Kontakt)

Veranstalter: Institut AIFB, Forschungsgruppe Wissensmanagement

Zu diesem Vortrag lädt das Institut für Angewandte Informatik und Formale Beschreibungsverfahren alle Interessierten herzlich ein.

A. Oberweis, H. Schmeck, R. Studer (Org.), Y. Sure-Vetter

Besucheranschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
Kaiserstr. 89
76133 Karlsruhe

Postanschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
76128 Karlsruhe