Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Marzia Cutajar		
Název práce	RDF Data Querying in Multi-Model NoSQL Databases		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Softwarové a datové inženýrství

Autor posudku	Jakub Klímek	**Role**	oponent
Pracoviště	KSI MFF UK		

Text posudku:

The theses addresses a relevant and current topic of alternative approaches to storage of RDF (graph) data and its querying. It is well structured, grammatically and typographically correct, and the related work was researched properly.

The thesis has two main parts. First, an approach to storing RDF data in ArangoDB, a multi-model NoSQL database is described. Second, an approach to translating queries from SPARQL, a query language used to query RDF, to ArangoDB's native query language AQL is described. Descriptions of both approaches are very formal, consisting of a series of **153 definitions**. While the formalism is correct and structured well, this makes the thesis hard to read. One way of tackling this would be to keep the complete set of definitions in an annex, and create some more readable description of the approaches and experiments.

Nevertheless, the approaches are implemented and properly evaluated. Even though in the end they are not better than other existing ones performance-wise, they are a nice scientific contribution to the state of the art which can be further built upon.

From a software engineering point of view, the experimental implementation is available as open-source on GitHub. However, it could be better documented. For instance, requirements on the software, such as the required `Gradle` version, are missing. Also, the part implementing the SPARQL to AQL translation requires a running database and also executes the query. This behavior can only be changed in code, while it would make more sense to implement a command-line parametr to enable running the tool with no external dependencies just to get the query translation.

A question for the defense:

- In 3.1 (page 26) the author mentions a benchmark in related work where RDF storage in ArangoDB outperforms traditional triplestores, and at the same time identifies a need to investigate SPARQL to AQL translations. How was the benchmark performed if there was no SPARQL to AQL translations used? Were the results comparable?

Finally, there are some minor issues:

- The `Indexes` section on page 20 would deserve more structure according to the index types described
- Definition 2.15 $d_i.\_key \neq d_i.\_key$ should be $d_i.\_key \neq d_j.\_key$
- In 3.2 (page 27) the last sentence of the first paragraph seems not to relate to SPARQL and R2RML
- Related work - SPARQL to XQuery translation - while it is clear how the query is translated, it is not clear on top of what the query runs, i.e. how the XML data looks like
- Figure 7.1 is not referenced from the text
- In GitHub command examples, absolute paths to files coming from the student's computer are stated, obviously useless to another user. At the same time, test data is not provided in the repository.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 17. 01. 2020

Podpis: