Technische Universität Berlin

TU Berlin | Fachgebiet ODS | TEL 4-1 | Ernst-Reuter-Platz 7 | 10587 Berlin

Fakultät IV Elektrotechnik und Informatik
Institut für Telekommunikationssysteme

Open Distributed Systems

Prof. Dr. Manfred Hauswirth

manfred.hauswirth@tu-berlin.de

Sekretariat TEL 4-1
Ernst-Reuter-Platz 7
10587 Berlin

Telefon +49 (0)30 314-78422
Telefax +49 (0)30 314-78420

Semantic Web Science Association

Berlin, 28. June 2016

Recommendation of the PhD thesis "Efficient, Scalable, and Provenance-Aware Management of Linked Data" by Marcin Wylot for the Semantic Web Science Association Distinguished Dissertation Award.

The thesis of Marcin Wylot investigates the following questions:

1. Are existing data management approaches sufficient for RDF data in terms of efficiency? If not, can they be adapted or is it better to design storage and query mechanisms from scratch specifically addressing the needs of RDF data management?
2. How can RDF data be managed efficiently in the cloud, thus providing scalability and elastic data management in a service-oriented fashion? What are the necessary storage layouts and partitioning/clustering approaches? What are the best query processing strategies and what are the trade-offs?
3. How can provenance information be tracked efficiently at different levels of granularity and how does this impact on the storage layout and query processing strategies?
4. How can RDF databases effectively support queries that take into account provenance and how does this influence query processing strategies, specifically, whether query processing can benefit from provenance?

Marcin in his thesis introduces a new storage model based on "molecule clusters" and "template lists" which enable an efficient co-location strategy and data structure for RDF data, tailored algorithms for partitioning data in cloud environments and efficient distributed query processing. The design of the data structures is special in various aspects: The data structures take into account recurring patterns occurring in the RDF data both on the data (physical) and schema (logical) levels, combining them into what the author calls "physiological" partitioning – a combination of aspects of tuple-based partitioning and graph-based partitioning. This technique is then shown to have various beneficial properties in

the efficient processing of (distributed) queries. Several techniques for partitioning in this general context are evaluated to identify an optimal one, applicable to most settings (adaptive partitioning). Those approaches are again extensively evaluated through large-scale experiments which show that DiploCloud can be orders of magnitude faster than state-of-the-art approaches on standard workloads.

Following, Marcin addresses the question of efficiently tracking provenance data practically for distributed RDF datasets. This is a very important but not well researched area as most research on provenance provides only theoretical approaches while engineered and tested solutions fit for purpose in practical deployments hardly exist. The author presents two storage models to physically co-locate lineage and instance data, and describes algorithms for tracing provenance at two granularity levels. The approaches are included in the TripleProv RDF store which is based on DiploCloud. The approaches are again evaluated with large-scale experiments which show that the overhead of including provenance is considerable. Yet the inclusion is necessary to establish the quality and trustworthiness of results. The developed approaches – though tested with DiploCloud – are generally applicable also to other RDF database systems with small modifications.

Finally, Marcin shifts the perspective from "only" tracing provenance to efficiently executing provenance-enabled queries over RDF data. A number of different query execution strategies for RDF queries which exploit provenance data are devised based on the approaches developed for TripleProv. These approaches are then again evaluated experimentally to find out the optimal strategies for various workloads. An interesting result from the research and evaluations presented is, that the use of provenance data can speed up query processing significantly due to the selectivity of provenance information, thus outweighing the costs of managing provenance data.

In summary, the thesis of Marcin Wylot has identified, analyzed and solved a significant scientific problem in the domain of RDF data management and provenance with the required scientific rigor. As discussed above, the contributions of this thesis to data management, query processing, and provenance in RDF databases are substantial and provide new scientific insight. The thesis is based on multi-authored papers, but the contributions of the candidate are clear from previous interactions and detailed discussions of the presented research at conferences and meetings of this evaluator with the candidate. The thesis provides a very systematic analysis of the problems at hand and then addresses all identified problems in a methodologically excellent way. A specific point in favour of the high quality of the thesis is that the candidate has not only developed the theoretical concepts but has underpinned the quality of the devised approaches by extensive, large-scale experiments, for which the descriptions, algorithms, implementations and the experimental data are available to enable repeatability of the experiments and comparison with new approaches to come. This is very good scientific practise. The relevance and contribution of the work is also already demonstrated by several publications in top-class international venues (WWW, ISWC, ESWC, TKDE). The thesis is very well written and easy to understand. Based on the above evaluation fully recommend this thesis to the SWSA Distinguished Dissertation Award.

Best regards,

(Prof. Dr. Manfred Hauswirth)