NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

CS SEMINAR

Title: **Scalable SPARQL Querying using Path Partitioning**

Speaker: Associate Professor Yongluan Zhou
Department of Mathematics and Computer Science
University of Southern Denmark

Date/Time: 13 August 2015, Thursday, 02:00 PM to 03:30 PM

Venue: MR1, COM1-03-19

Chaired by: Dr Tan Kian Lee, Shaw Senior Professor, School of Computing
(tankl@comp.nus.edu.sg)

Abstract:

The emerging need for conducting complex analysis over big RDF datasets calls for scale-out solutions to process big RDF datasets. Queries over RDF data often involve complex self-joins, which would be very expensive to run if the data are not carefully partitioned across the cluster and hence distributed joins over massive amount of data are necessary.

Existing RDF data partitioning methods can nicely localize simple queries but still need to resort to expensive distributed joins for more complex queries. In this paper, we propose a new data partitioning approach that makes use of the rich structural information in RDF datasets and minimizes the amount of data that have to be joined across different computing nodes. We conduct an extensive experimental study using two popular RDF benchmark data and one real RDF dataset that contain up to billions of RDF triples. The results indicate that our approach can produce a balanced and low redundant data partitioning scheme that can avoid or largely reduce the cost of distributed joins even for very complicated queries.

In terms of query execution time, our approach can outperform the state-of-the-art methods by orders of magnitude.

Biodata:

Yongluan Zhou is an Associate Professor in the Department of Mathematics and Computer Science at SDU. Before joining SDU, he had been working at EPFL as a postdoctoral researcher. He obtained his PhD in Computer Science at National University of Singapore. His research interests span across database systems and distributed systems, especially in the area of query processing and optimization and data stream management.