

Querying big data just got universal

June 24 2019



(L-R) Fuad Jamour, Panos Kalnis and Yanzhao Chen are building systems and algorithms for processing and analyzing very large datasets. Credit: 2019 KAUST

To solve one of the key obstacles in big-data science, KAUST researchers have created a framework for searching very large datasets that runs easily on different computing architectures. Their achievement allows researchers to concentrate on advancing the search engine, or query engine, itself rather than on painstakingly coding for specific computing platforms.

Big data is one of the most promising yet challenging aspects of today's information-heavy world. While the huge and ever-expanding sets of information, such as online-collected data or [genetic information](#), could hold powerful insights for science and humanity, processing and interrogating all this data require highly sophisticated techniques.

Many [different approaches](#) to querying [big data](#) have been explored. But one of the most powerful and computationally effective is based on analyzing data with a subject-predicate-object triplestore structure of the form (e.g., apple, is a, fruit). This structure lends itself to being treated like a graph with edges and vertices, and this characteristic has been used to code query engines for specific computing architectures for maximum efficiency. However, such architecture-specific approaches cannot be readily ported to different platforms, limiting the opportunities for innovation and advancement in analytics.

"Modern computing systems provide diverse platforms and accelerators, and programming them can be intimidating and time consuming," say Fuad Jamour and Yanzhao Chen, Ph.D. candidates in Panos Kalnis's group in KAUST's Extreme Computing Research Center. "Our research group focuses on building systems and algorithms for processing and analyzing very [large datasets](#). This research addresses the desire to write a program once and then use it across different platforms."

Rather than the previously used graph-traversal or exhaustive relational-indexing approaches, the group queried triplestore data by using an applied mathematical approach called sparse-matrix algebra.

"Our paper describes the first research graph-query [engine](#) with matrix algebra at its core to address the issue of portability," says Jamour. "Most existing graph-query engines are designed for single computers or small distributed-memory systems. And porting existing engines to large distributed-memory systems, like supercomputers, involves significant

engineering effort. Our sparse-matrix algebra scheme can be used to build scalable, portable and efficient graph-query engines."

The team's experiments on large-scale real and synthetic datasets achieved performance comparable with, or better than, existing specialized approaches for complex queries. Their scheme also has the capacity to scale up to very large computing infrastructures handling datasets of up to 512 billion triples.

"These ideas can facilitate building analytics components in graph databases with cutting-edge performance, which is currently in high demand," says Chen.

More information: Fuad Jamour et al, Matrix Algebra Framework for Portable, Scalable and Efficient Query Engines for RDF Graphs, *Proceedings of the Fourteenth EuroSys Conference 2019 CD-ROM on ZZZ - EuroSys '19* (2019). [DOI: 10.1145/3302424.3303962](https://doi.org/10.1145/3302424.3303962)

Provided by King Abdullah University of Science and Technology

Citation: Querying big data just got universal (2019, June 24) retrieved 2 May 2024 from <https://phys.org/news/2019-06-querying-big-universal.html>

<p>This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.</p>
--