New data-mining strategy that offers unprecedented pattern search speed could glean new insights from massive datasets

December 28 2016

A new data mining strategy that offers unprecedented pattern search speed could lead to new insights from massive data. Credit: © Mopic / Alamy Stock Photo DTFTEM

Searching for recurring patterns in network systems has become a fundamental part of research and discovery in fields as diverse as

biology and social media. KAUST researchers have developed a pattern or graph-mining framework that promises to significantly speed up searches on massive network data sets.

"A graph is a data structure that models complex relationships among objects," explained Panagiotis Kalnis, leader of the research team from the KAUST Extreme Computing Research Center. "Graphs are widely used in many modern applications, including social networks, biological networks like protein-to-protein interactions, and communication networks like the internet."

In these applications, one of the most important operations is the process of finding recurring graphs that reveal how objects tend to connect to each other. The process, which is called frequent subgraph mining (FSM), is an essential building block of many knowledge extraction techniques in social studies, bioinformatics and image processing, as well as in security and fraud detection. However, graphs may contain hundreds of millions of objects and billions of relationships, which means that extracting recurring patterns places huge demands on time and computing resources.

"In essence, if we can provide a better algorithm, all the applications that depend on FSM will be able to perform deeper analysis on larger data in less time," Kalnis noted.

Kalnis and his colleagues developed a system called ScaleMine that offers a ten-fold acceleration compared with existing methods.

"FSM involves a vast number of graph operations, each of which is computationally expensive, so the only practical way to support FSM in large graphs is by massively parallel computation," he said.

In parallel computing, the graph search is divided into multiple tasks and

each is run simultaneously on its own processor. If the tasks are too large, the entire search is held up by waiting for the slowest task to complete; if the tasks are too small, the extra communication needed to coordinate the parallelization becomes a significant additional computational load.

Kalnis' team overcame this limitation by performing the search in two steps: a first approximation step to determine the search space and the optimal division of tasks and a second computational step in which large tasks are split dynamically into the optimal number of subtasks. This resulted in search speeds up to ten times faster than previously possible.

"Hopefully this performance improvement will enable deeper and more accurate analysis of large graph data and the extraction of new knowledge," Kalnis said.

Provided by King Abdullah University of Science and Technology