Name: Xifeng Yan

Affiliation: Computer Science Department, University of California at Santa Barbara

Email: xyan@cs.ucsb.edu

Two Topics:

3. Data Portability, Consistency, and Management
8. Cloud Self-Monitoring and Autonomic Control

Why not add a topic for improving Cloud Computing with Data Mining techniques

Summary of My Current Research Activities Related to Cloud Computing:

Searching and mining large graphs today is critical to a variety of application domains, ranging from community detection in social networks and blog spaces, to searches for functional modules in biological pathways. These tasks often involve complex queries that repeatedly access huge amounts of graph links, exposing issues unexplored in traditional keyword queries. The inter-connected nature of graphs mean their operations tend to crawl across many links, resulting in extremely large memory footprints that stretch the capabilities of today's commodity servers. Massive graphs must be carefully partitioned and distributed across clusters to avoid costly inter-node communication that increases latency and serializes operations between nodes. Such partitions also must adapt in real time to topology changes in both the graph and cluster configuration

We have developed a distributed graph processing system to serve various graph queries with a set of scalable algorithms to monitor/analyze query workload changes, and dynamically repartition graphs with regard to these changes. It is an effort to systematically develop a cloud computing infrastructure that provides high level abstractions for graph primitives, simplifying design of complex queries while addressing difficult challenges of maximizing data parallelism and adaptive graph partitioning across clusters.

We have developed two real applications, correlation analysis in information network and de novo short read assembly in de Bruijn graphs, to test the performance of the proposed system.

Future Research: Data Mining Techniques for Managing and Diagnosing Cloud Computing Systems

Cloud computing promises high scalability, flexibility and cost-effectiveness to satisfy emerging computing requirements. To efficiently provision computing resources in the cloud, system administrators need the capabilities of analyzing server workload and diagnosing software dependency.

It is possible to search repeatable workload patterns by exploring cross-server correlations resulted from the dependencies among applications running on different servers. Treating server workload data samples as multiple time series, we are testing a co-clustering technique to identify groups of servers that frequently exhibit correlated workload patterns, and also the time periods in which these server groups are active. The results could not only help system administrators better understand group-level workload characteristics in a cloud, but also make more accurate predictions on workload changes over time.

We can further deepen this study by mining software dependency in Cloud Computing. Such dependency will help us better diagnose complicated software problems arising from Cloud Computing, including configuration errors, software inconsistency, etc.

We believe data mining techniques could significantly improve the service quality and automate the management of Cloud Computing systems.