Scalable Retrieval Techniques for Massive High-Dimensional Data (PI: Dr. Yiu Man Lung; 2014/15)

Similarity search has a broad range of applications such as near-duplicate detection systems (TinEye), public video surveillance systems, photo sharing sites (Flickr), and sensor data analysis. In these applications, data objects are represented as high-dimensional points, and users wish to find data objects similar to a specified query object. These applications are facing two key challenges nowadays: (i) managing a massive amount of data objects, and (ii) processing queries at a rapid rate. The massive data amount is due to widespread usage of devices (e.g., cameras, sensors), whereas the rapid query rate is caused by the vast number of users or monitoring targets. These challenges hinder the above applications from reporting accurate query results to a large number of users in reasonable response time. Although there has been extensive academic research on efficient similarity search, existing methods fail to tackle the challenges (i) and (ii). These methods assume that a single machine can accommodate all data and indexes, which no longer hold for the above applications. A promising approach is to adopt a distributed architecture with multiple commodity machines for storing data, indexes, and processing queries. This approach scales well with the data size and the query throughput, by deploying more machines. In this project, we study how to improve the query throughput of this system from the data engineering

perspective, at a fixed number of machines. This project aims at developing techniques to optimize the query throughput of similarity search on massive high-dimensional data in a distributed architecture. Our track records in multi-dimensional data processing and distributed databases are essential to the success of this project. The performance bottlenecks are caused by the network transfer cost and the disk access cost. To achieve high query throughput, we will investigate three fundamental directions towards alleviating the above bottlenecks: (i) partitioning, (ii) caching, and (iii) reducing disk access time for high-dimensional data. We will carefully exploit the characteristics and distribution of high-dimensional data in designing optimization techniques. Also, we will build (iv) a prototype similarity search system for massive high-dimensional data with our proposed techniques, and release it as open-source software. We envision that this project will advance the query processing techniques for massive high-dimensional data, and improve the query throughput significantly for a broad range of applications, e.g., near-duplicate detection systems, public video surveillance systems, photo sharing sites, and sensor data analysis.