Abstract

Clustering and nearest neighbor searches in high-dimensions are fundamental components of computational geometry, computational statistics, and pattern recognition. Although there is a significant body of work for efficient clustering and nearest neighbor searches in database systems, there is surprisingly little work on algorithms and libraries for high-end scientific computing platforms. In particular, libraries and algorithms that allow simulation codes to be coupled directly with data analysis for end-to-end parallelism and scale to thousands of cores are absent.

We present algorithms built on top of Message Passing Interface (MPI) and OpenMP. These algorithms enable the solution of large-scale computational geometry problems on massively parallel computer architectures. In particular, we combine parallel distance and filtering operations and kmeans clustering algorithms with an optimized seeding procedure, locality-sensitive hashing, and a novel parallel indexing tree data structure to support exact and approximate basic operations for problems in high-dimensional computational geometry.

We outline the algorithms and provide preliminary performance benchmarks test. In our tests, the overall scheme shows excellent scalability and enables the analysis of datasets of unprecedented scale. In our largest runs, we were able to conduct searches and clustering on datasets with over four billion reference points in 1000 dimensions (a 45TB dataset) in under 30 seconds on 98K cores on the NSF NICS Kraken platform. Tree construction takes less than three seconds for a dataset with over 200 million points in 100 dimensions on 12288 cores (2048 MPI processes).