Multidimensional Adaptive and Progressive Indexes

Matheus Agio Nerone[1], Pedro Holanda[1], Eduardo C. De Almeida[2], Stefan Manegold[1]

[1] Centrum Wiskunde & Informatica (CWI)
[2] Federal University of Paraná (UFPR)

Exploratory data analysis is used by data scientists to extract knowledge from new data. The analysis consists of an interactive trial and error process, where the scientist depends on previous results to formulate new hypotheses. Because the analysis is naturally interactive, the system needs to keep low response times. Otherwise, the scientist can lose focus. These analyses ordinarily consist of workloads with highly selective multidimensional queries, which need multidimensional indexes for efficient searches. However, creating such indexes is a challenging problem. Because of the exploratory nature, previously acquired knowledge may end up not being valid anymore. Furthermore, creating all possible indexes is also impracticable because of the vast number of possibilities.

In this work, we identify four objectives that are desirable for exploratory data analysis: (1) low overhead over the initial queries, (2) low query variance (i.e., high robustness), (3) predictable index convergence, and (4) low total workload time. We also propose three novel incremental indexing techniques: (a) The Adaptive KD-Tree has the lowest total workload time at the expense of a higher indexing penalty for the initial queries, lack of robustness, and unpredictable convergence. (b) The Progressive KD-Tree has predictable convergence and a user-defined indexing cost for the initial queries. However, total workload time can be higher than with Adaptive KD-Trees, and per-query time still varies. (c) Greedy Progressive KD-Tree aims at being more robust at the expense of only improving the per-query cost after full index convergence.

Our experimental evaluation using both synthetic and real-life data sets and workloads shows that: (a) the Adaptive KD-Tree can reduce total workload time by up to a factor 2 compared to the state-of-the-art, (b) the Progressive KD-Tree achieves predictable convergence with up to one order of magnitude lower initial query cost, and (c) the Greedy Progressive KD-Tree is the more robust, exhibiting the lowest query variance up to three orders of magnitude lower than the state-of-the-art.