CS561 Spring 2022 - Research Project

Title: WOODBLOCK in Qd-tree

Background: To reduce the response time for real-time analytical, data partitioning is a promising technique to avoid unnecessary data retrieving. For example, two traditional techniques in state-of-the-art database systems (e.g., MySQL [1]) are range and hash partitioning. However, these two strategies usually cannot best fit the query workload. Hence, a workload-aware partitioning strategy is required to adapt the given query set. On the other hand, optimally assigning records into blocks/pages is an NP-Hard problem [2]. WOODBLOCK [3] is a typical technique that applies deep Reinforcement Learning to split the high-dimension space, using extracted predicates from history query workload.

Objective: The overall goal of our project is to implement WOODBLOCK to construct the query-data routing tree based on a given dataset and a query workload.

(a) Step 1: Get familiar with block searching strategies in "Fine-grained partitioning for aggressive data skipping." [2], WOODBLOCK descriptions in "Qd-tree: Learning data layouts for big data analytics."[3]
(b) Step 2: Get familiar with NeuralCuts [4] (which the WOODBLOCK is implemented based on). Based on the WOODBLOCK description in Qd-tree, modify states, actions, rewards definitions in NeuralCuts code.
(c) Step 3: Parse the emulated queries and dataset into some format that can be used in WOODBLOCK. Estimate the I/O cost using the produced Qd-tree
(d) Step 4: Benchmark WOODBLOCK with different queries and dataset (varying the selectivity, the skewness, the data dimension and the noisy queries in testing set)

Responsible mentor: *Zichen Zhu*

References

[1] MySQL Partitioning Types, Oracle Corporation and/or its affiliates, https://dev.mysql.com/doc/refman/8.0/en/partitioning-types.html
[2] Sun, Liwen, et al. "Fine-grained partitioning for aggressive data skipping." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. 2014.
[3] Yang, Zongheng, et al. "Qd-tree: Learning data layouts for big data analytics." Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 2020.
[5] NeuroCut, https://github.com/neurocuts/neurocuts