HW2

Yuting Wang yw3167

Operator Reordering:

We try to move more selective operators upstream to filter data early.

Data:

The data is a randomly generated dataset including over 30000 data points between (1,500).

Performance Measurement:
We measure the performance as throughput versus selectivity of B in both reordered or not reordered situations.

Suppose we have two operators A and B, they have the equal cost. The selectivity of A is 0.5, selectivity of B changes from zero to one.

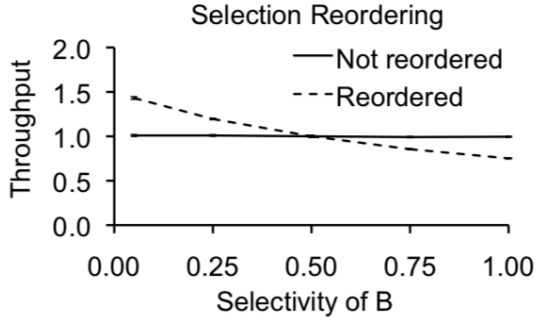
From A to B:
Data is first processed by A, then A will process all the data and B will process half of the input
Throughput from A to B is proportional to

$$\frac{1}{cost(A) + s(A) * cost(B)}$$

From B to A:
Data is totally processed by B firstly, while the amount of data that A is going to perform is relevant to the selectivity of B. The final throughput will be higher if B has a smaller selectivity.
Throughput from B to A is proportional to

$$\frac{1}{cost(B) + s(B) * cost(A)}$$

The ideal throughput-selectivity relationship is shown below:



Results:

Normalization:

We normalized the throughput of A-to-B to 1 and plot $\frac{\text{throughput(B to A)}}{\text{throughput(A to B)}}$ as the throughput in the graph. The final result is shown below, which includes a fluctuation. But the overall trend of throughput is from 1.5 to 0.75.