Paper Review

“ Output-optimal Parallel Algorithms for Similarity Joins ”

1. Summary

Parallel join, which is used for table A join to table B, have received much attention. This paper was improved from Beame’s work. The output size in reality is smaller than many papers’ assumptions. This paper first improve Beame’s algorithm to true optimality.

Similarly join problem is let A join B where distance (A, B) is less than a value. The distance function can be any.

The previous work’s assumptions may be hard to satisfy in reality though they can get good performance in some conditions.

In MPC models, there are some primitives, sorting, multi-numbering, sum-by-key, multi-search, cartesian product and server allocation.

The EQUI-JOIN algorithm is first compute $OUT = \Sigma_v N_1(v) N_2(v)$ by sum-by-key and add up. Then, compute the join.

The algorithm of JOIN UNDER l_2. First, this paper construct a partition tree, then partially compute covered cells, finally compute fully covered cells.

This paper include so much mathematic formulas and it’s difficult for me to fully understand it.

2. Advantages

+ This paper archive great complexity when the size of tuple is relatively smaller, which meet more real requirements. This paper gets more practicality.

3. Disadvantages

– There exists another arguments except OUT.

QIN FEIRAN qinfr@shanghaitech.edu.cn 21/04/2022