Database Systems - project 2 report

Lucie Hoffmann

19 May 2021

1 Evaluation parameters

For the performance and accuracy evaluation, we set the seed to be 43, the number of partitions for BaseConstructionBalanced to be 16 and the threshold for ExactNN to be 0.6.

2 Performance comparison

In Figure 1 we see the difference in runtime between ExactNN, BaseConstruction, BaseConstructionBroadcast and BaseConstructionBalanced for query sets 1-2 skew and not skew. Interestingly, we see that ExactNN is the second fastest implementation after BaseConstruction. All implementations have more or less the same performances for both skew and not skew.

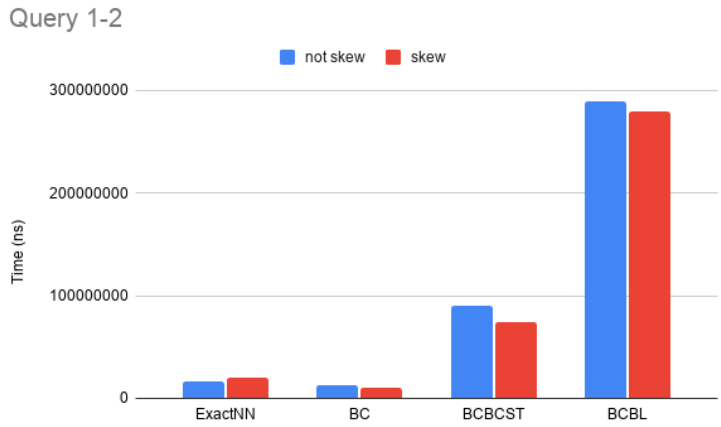

Figure 1: Difference in runtime between constructions with small-size query sets with few duplicates (one skew and the other not).

In Figure 2 we see the difference in runtime for query sets 1-10 skew and not skew. In general, runtime is better than for the previous query sets, prob-

ably due to the higher amount of duplicates. The overall comparison between implementations is still the same.

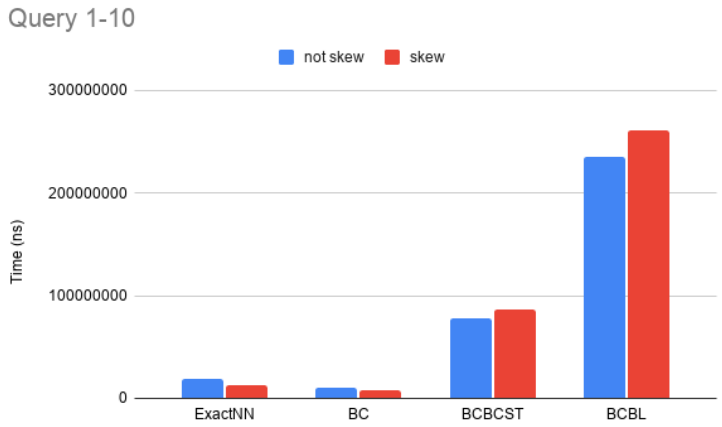

Figure 2: Difference in runtime between constructions with small-size query sets with more duplicates (one skew and the other not).

In Figure 3 we see the difference in runtime for query sets 10-2 skew and not skew. As the number of queries and corpus are bigger than previously, the difference between skew and not skew is more obvious for BaseConstruction-Broadcast and BaseConstructionBalanced.

In Figure 4 we see the difference in runtime for query sets 10-10 skew and not skew. The difference between skew and not skew is not so obvious as the number of duplicates is higher.

In Figure 5 we see the difference in runtime for query sets 20-2 skew and not skew. Here the difference between skew an not skew is even more obvious for BaseConstructionBalanced.

In Figure 6 we see the difference in runtime for query sets 20-10 skew and not skew. Similarly, we see a big difference between skew and not skew for BaseConstructionBalanced, but this time skew is worse.

The bigger the amount of data, the higher the processing time and the greater the difference between skew and not skew. BaseConstructionBroadcast and BaseConstructionBalanced are more hurt by the increase in data and BaseConstructionBalanced is more hurt by the difference in skew. In every case, BaseConstruction is always the implementation with the best runtime.

3 Accuracy comparison

In Figure 7, we see the difference in average distances over all corpuses, for queries with and without skew and few duplicates.

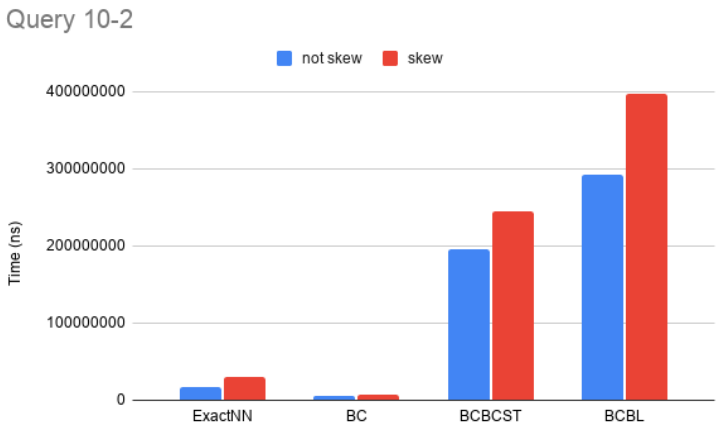

Figure 3: Difference in runtime between constructions with medium-size query sets with few duplicates (one skew and the other not).

In Figure 8, we see the difference in average distances over all corpuses, for queries with and without skew and more duplicates.

Interestingly, the average distance of ExactNN is not always the best. This may be due to the choice of the threshold. We see in general that the bigger the amount of data, the worse the distance, especially for skew data.

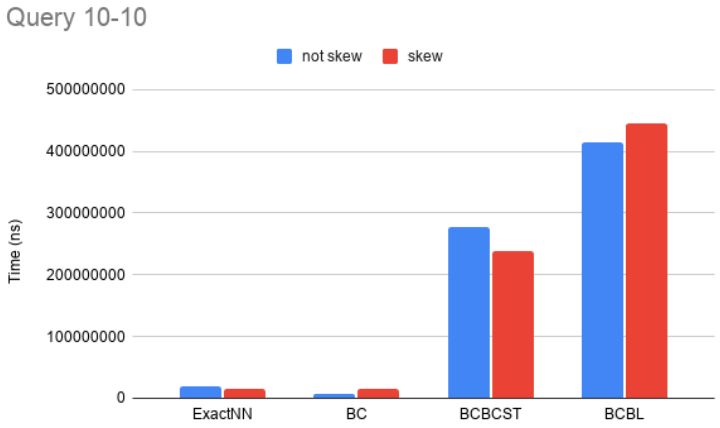

Figure 4: Difference in runtime between constructions with medium-size query sets with more duplicates (one skew and the other not).

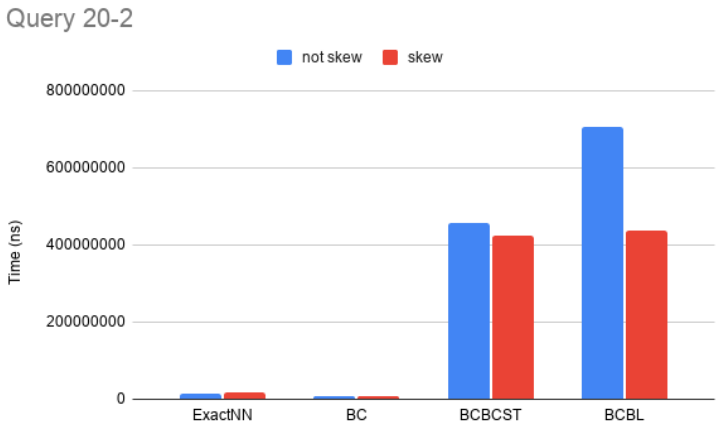

Figure 5: Difference in runtime between constructions with big-size query sets with few duplicates (one skew and the other not).

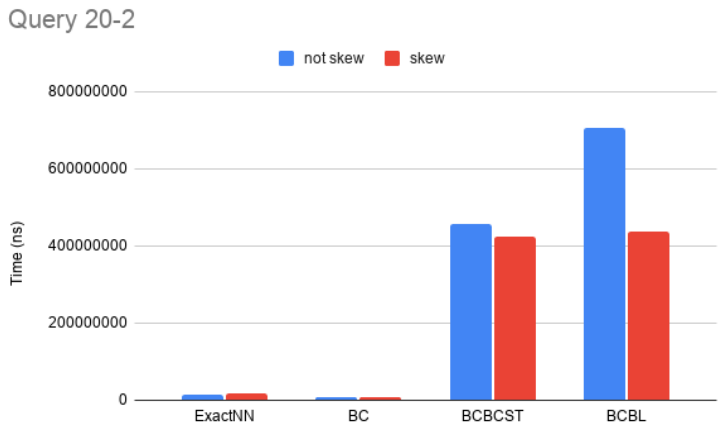

Figure 6: Difference in runtime between constructions with big-size query sets with more duplicates (one skew and the other not).

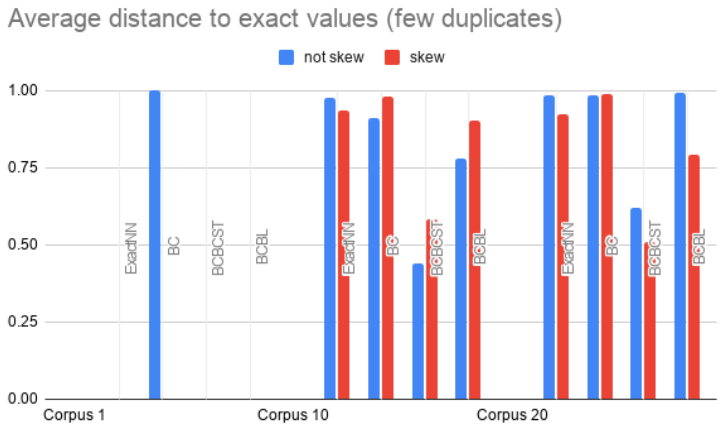

Figure 7: Difference in average distance between constructions with 3 different corpuses of increasing size with few duplicates (one skew and the other not).

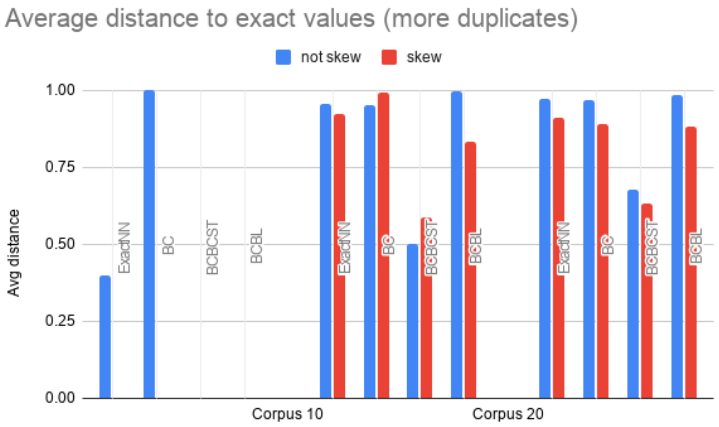

Figure 8: Difference in average distance between constructions with 3 different corpuses of increasing size with more duplicates (one skew and the other not).