Work Progress

kNN Search with Parallel Incremental Query Answering

Jaouhara Chanchaf

Tuesday Feb 1st, 2023

1. Summary

Done:

AI 0	Find a spark cluster to run JOSIE experiments.
AI 1	Kashif: Only sort NNs that were not returned (incrementally) and check for performance improvement.
AI 3	Check memory usage for Kashif.

In progress:

AI 2	Search for alternative data structures to store and process kNNs and explore their complexity.

2. Kashif: Insert NN time vs Sort time

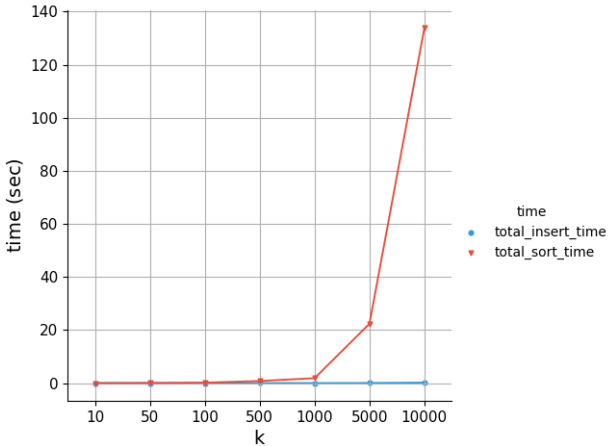

Figure 1: Kashif total NN insert and sort
time(1 queries, query size = 100, dataset = 100k tables, 490k cols, 5M vectors)

3. Kashif: Only sort running NNs

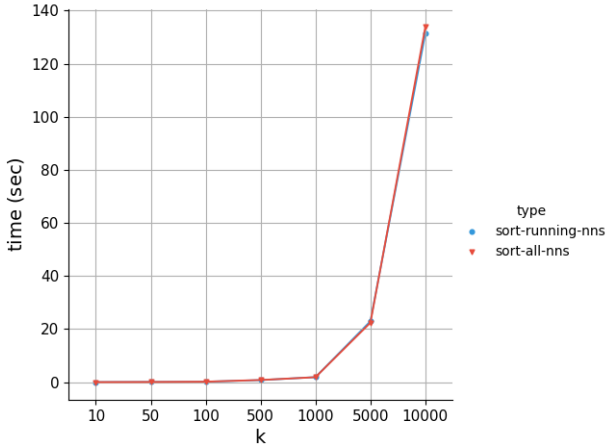

Figure 2: Kashif impact of only sorting running NNs on sort time
(1 queries, query size = 100, dataset = 100k tables, 490k cols, 5M vectors)

→ Only sorting running NNs does not reduce the sort time significantly.

4. Discussion

- We need the sorted array structure to :
  1) Get k-th NN (max distance). 2) Return incremental results.
- The current implementation does not take advantage of the already sorted array when inserting a new element. The insert operation has an average time complexity of $O(n^2)$.
- Use Min-Max Heap for storing kNNs?

4. Discussion

- We need the sorted array structure to :
  1) Get k-th NN (max distance). 2) Return incremental results.
- The current implementation does not take advantage of the already sorted array when inserting a new element. The insert operation has an average time complexity of $O(n^2)$.
- Use Min-Max Heap for storing kNNs?

Worst Case Time Complexity:

	Sorted Array	Min-Max Heap
Insert(d)	$O(n^2)$	$O(0.5 * log(n+1))$
GetMin()	$O(1)$	$O(1)$
GetMax()	$O(1)$	$O(1)$
DeleteMax()/ DeleteMin()	$O(1)$	$O(2.5 * log(n))$

In Kashif algorithm : InsertNN(d) = DeleteMax() + Insert(d)

5. Kashif: Insert NN, get i-th NN and get k-sth NN count

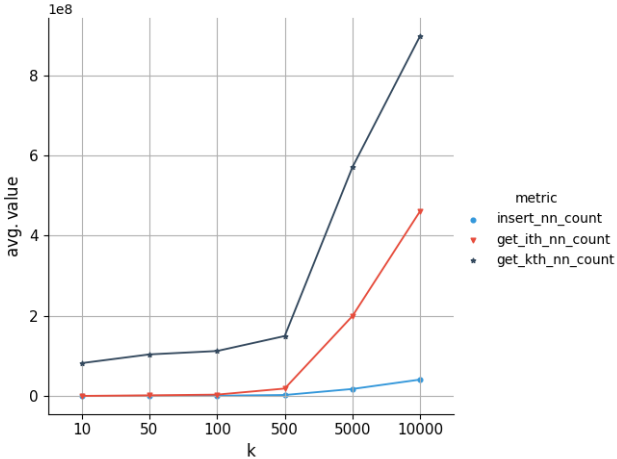

Figure 3: Kashif Average Count #insertNN, #get-ithNN and #get-kthNN
(10 queries, query size = 100, dataset = 100k tables, 490k cols, 5M vectors)

5. Kashif: Insert NN, get i-th NN and get k-sth NN count

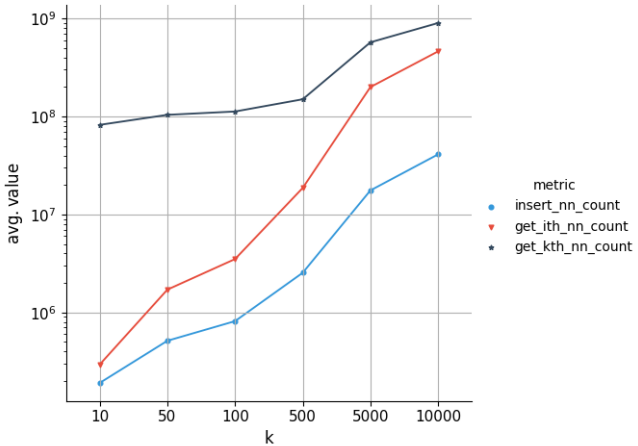

Figure 4: Kashif Average Count #insertNN, #get-ithNN and #get-kthNN (log scale) (10 queries, query size = 100, dataset = 100k tables, 490k cols, 5M vectors)

5. In progress

- Tested C++ implementation of Min-Max Heap: `https://github.com/skarupke/heap`
- Implement Min-Max Heap in C.