Work Progress

kNN Search with Parallel Incremental Query Answering

Jaouhara Chanchaf

Tuesday Nov. 1st, 2022

1. Summary

Done:

AI 1	Kashif: Measure the query time for Parallel Incremental Query Answering with multiple worker-threads.
AI 2	Read Hydra 2

In progress:

AI 2	Read Progressive SS with Probabilistic Quality Guarantees
AI 3	Find a technique to estimate recall.

Issues:

- Cannot create index for larger datasets (5 Million tables).

1. Performance Comparison

Experiment over 1M tables, 4.9M columns, 50.3M vectors (9 GB).

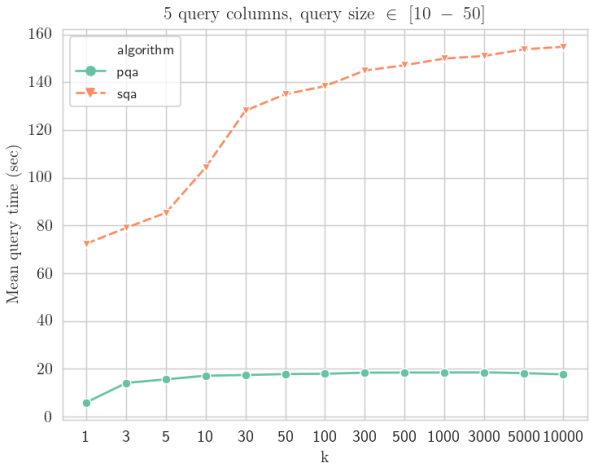

Figure: Kashif performance: Parallel (Incremental) Query Answering (pqa) Vs Sequetial Query Answering (sqa)

(!) Waiting for brute force results to measure recall and precision.

2. Progressive kNN search

Data Series Progressive Similarity Search with Probabilistic Quality Guarantees (A. Gogolou SIGMOID, 2020):

- **What:** How to perform Progressive NN search with Probabilistic Guarantees?
- **Why:** To insure interactive response time when performing search over large collections of data series.
- **How:** A probabilistic method that provides quality guarantees for intermediate results.

2. Progressive kNN search

Data Series Progressive Similarity Search with Probabilistic Quality Guarantees (A. Gogolou SIGMOID, 2020):

1. Good answers are found at the early stages of kNN search.
2. Return results progressively.
3. Early results can lead users to false conclusions.
4. Provide users with probabilistic guarantees on the quality of progressive results:
   - How close is the current result to the exact answer?
   - The probability that the current answer is the exact answer.
   - When it is expected to find the exact answer?

4. Discussion

- Incremental query answering is not timed. The user does not know when the next incremental answer will be returned.
- In incremental query answering the quality of returned results is guaranteed by the the overall recall is not.
- What if k_{max} is underestimated, low recall?
- Approximating the pdf using the kNN distance distribution of witnesses requires defining k in the training phase (to estimate the 1nn, 2nn, ..., knn distance distributions).