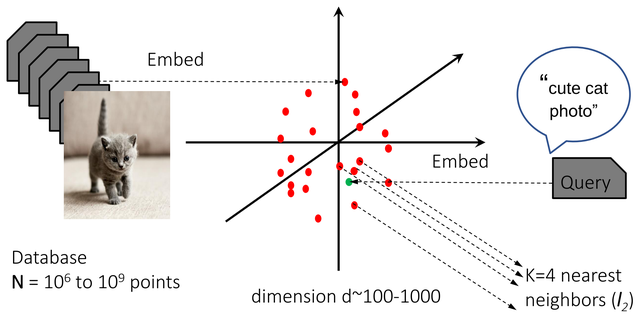
Investigation on Sim search data set via ParlayANN library

Taekseung Kim, Guy Blelloch, Magdalen Dobson *(Carnegie Mellon University)*

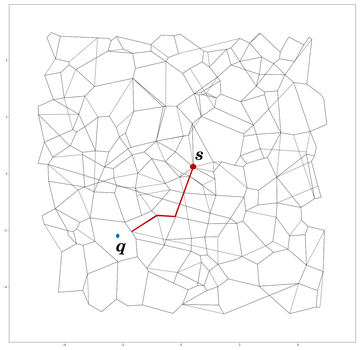
Introduction

K Nearest Neighbor search problem



Current Solution

Graph based ANN indices



Index: One vertex per data point/embedding. Directed edges between vertices.

For query *q*:
Start at point ***s***, and iterate:
1. compute dist. from ***q*** to neighbors
2. hop to node closest to ***q***, *as long as distance improves*

Index build: Create a low-degree graph that guides queries with the **fewest hops** and **distance comparisons**

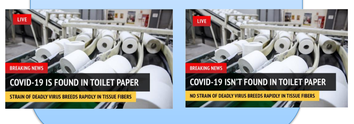
Starting from one vertex per data

Problem: Graph Based algorithm is the fastest solution, but performs bad on specific dataset

BIGANN dataset

1 billion size vector,128 dimension

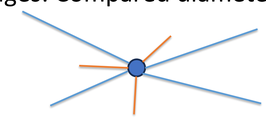
SimSearchNet dataset
-Near duplicate vector



Methodology

Hypothesis 1: **Near-duplicates** cause **cluster-like structures** in the graph and thus prevent the greedy search from converging

1. Found the **distribution of the short edges** of generated graph in SimSearchNet dataset.
   Short edges: Compared diameter of a dataset. With more short edges, there likely is a cluster.



Number of short edges: 3

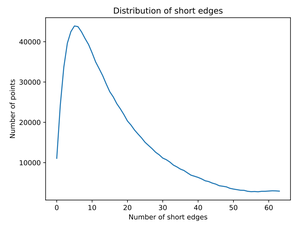
2. Check the **distance distribution** of SimSearchNet dataset, compared to BIGANN dataset

Hypothesis 2: Starting from **multiple points** would make the algorithm faster.

3. Implement a new method of search. Start from **two starting points**, using the second starting point as the true closest neighbor.

Results

Experiment 1: Distribution of short edges

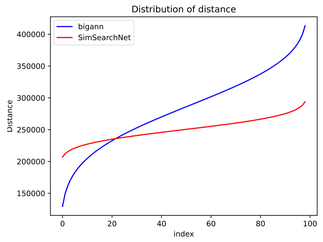


We found there are not a lot of short edges among the vertices

Rather, most of the points have 0~6 short edges, where their maximum degree is 64

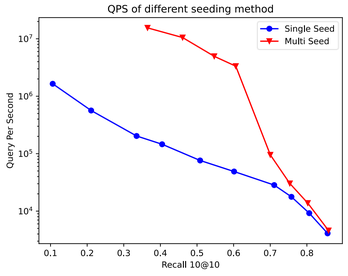
This fact implies there might not be a lot of clusters

Experiment 2: Distance distribution of the SimSearchNet dataset, compared to BIGANN dataset



SimSearchNet dataset has a narrower distance distribution, which means the distance among points is similar. This prevents the greedy search from converging.

Experiment 3: QPS(Query Per Second) comparison of multiple and single starting point

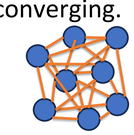


We can see that the multi seed method is much faster at finding 0.6 recall(which is 6 correct points out of 10), but gets almost the same throughput(Although it is about 1.4 times faster)

Discussion

Hypothesis 1:

- **Number of Short edges:** There are **not a lot of clusters** in SimSearchNet dataset, since there is small amount of short edges
- **Distance distribution**: Compared to the BIGANN dataset, SimSearchNet dataset has a **narrow distribution**. A narrow distribution is likely to prevent the greedy search from converging.



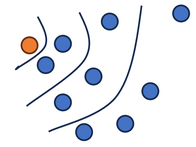
One cluster induces vast amount of short edges

Hypothesis 2:

- **Seeding method:** Searching from multiple points, we found that by providing one nearest neighbor we get 4 of the nearest points immediately. However, the **overall speed** for the 10 nearest neighbors has x**1.4 speedup**, which is not as much as we expected.

Future Work

- We plan to try more **optimizations** on searching and try some other searching algorithms.
- **Other prospective methods: Bucket-based methods** can be integrated into the search algorithm to speed up search for the SimSearchNet dataset.



3 buckets, based on distance of source

Contact
taekseuk@andrew.cmu.edu