YANQI CHEN

12011319@mail.sustech.edu.cn · (+86) 18372058013 · whateveraname

EDUCATION

Southern University of Science and Technology **Shenzhen, China**
***Bachelor** in Computer Science and Engineering (CSE)* Aug. 2020 – (Expected) June. 2024
GPA: 3.86/4.00, **Ranking**: 11/220
Core Courses: Linear Algebra (4.0 / 4.0), Probability and Statistics (3.94 / 4.0), Algorithm Design and Analysis (3.94 / 4.0), Digital Logic (3.94 / 4.0), C/C++ Program Design (3.94 / 4.0), Database Principles (3.94 / 4.0), Operating System (3.94 / 4.0), Artificial Intelligence (3.94 / 4.0)
Programming Languages: C/C++, Python, Java; **Tools**: Git, LaTeX, CMake, Docker

RESEARCH INTERESTS

Vector Search, Graph Processing, Database Systems, Information Systems, Data Mining

RESEARCH PROJECTS

Database Group, Southern University of Science and Technology

Research Assistant, Advisor: Prof. Xiao Yan Mar. 2022 – Oct. 2023

Approximate K-Nearest Neighbor Graph Construction
K-nearest neighbor graph (KNNG) connects each vector to its K-nearest neighbors and has many applications in data mining and machine learning. The project goal is to build KNNG efficiently for large datasets.

- For algorithm, initialize high quality neighbors for each vector with inverted index and dynamically adjust the parameters of NN-Descent (e.g. iteration time and neighbor sample size) to reduce execution time.
- For implementation, improve NN-Descent code, e.g., using SIMD instructions to accelerate distance computation and inplace candidate pool update to avoid unnecessary data copy.
- **Research Output**: SIGMOD'23 Programming Contest **World Finalist**

Approximate Nearest Neighbor Search (ANNS) for Out-of-Distribution (OOD) Queries
ANNS algorithms have severe performance degradation when the query distribution does not match the data distribution. The project goal is to design algorithms that work well for OOD queries.

- Improve graph-based index and propose to start graph traversal from multiple entry points identified by a K-means tree, which resolves the problem of graph connectivity and reduce the length of detours.
- Apply scalar quantization (SQ) to the database vectors to reduce memory traffic in distance computation.
- **Research Output**: NeurIPS'23 Big-ANN Competition OOD Track **3rd Place**

Data Curation Lab, Rutgers University

Research Assistant, Advisor: Prof. Dong Deng June. 2023 – Present

Billion-Scale Dataset Deduplication (currently ongoing)
Dataset deduplication removes duplicates from the dataset and is widely used in model training. The project goal is to design a framework to deduplicate billion-scale datasets with both high quality and good efficiency.

- Propose a novel deduplication approach, which first builds a range graph (where a vector is connected with vectors having a distance smaller than r) and then solves the Maximal Independent Set problem on it.
- Design a disk-based range graph construction procedure to handle large datasets. Tackle the disk I/O bottleneck problem by caching disk data in memory with Belady's cache replacement algorithm and reordering disk reads using Graph Ordering algorithm to maximize cache efficiency.

HONORS & AWARDS

3rd Place, NeurIPS'23 Big-ANN Competition OOD Track 2023
Finalist, SIGMOD'23 Programming Contest 2023
3rd Prize, Scholarship for Outstanding Students 2022
2nd Prize, Scholarship for Outstanding Students 2021