YANQI CHEN

yanqichen@umass.edu · (+1) 4138005660

EDUCATION

University of Massachusetts Amherst — **Amherst, MA, USA**
***PhD** in Computer Science* — Sep. 2024 – (Expected) Jun. 2029
Advisor: Alexandra Meliou

Southern University of Science and Technology — **Shenzhen, China**
***Bachelor** in Computer Science and Engineering (CSE)* — Sep. 2020 – Jun. 2024
GPA: 3.86/4.00, **Ranking**: 11/220

Skills: C/C++, Python, Java, Git, LaTeX, CMake, Docker

RESEARCH INTERESTS

Vector Search, Database Systems, Information Systems

RESEARCH PROJECTS

Large-scale Vector Similarity Join with SSD
Similarity join (SJ)—a widely used operation in data science—finds all pairs of items that have distance smaller than a threshold. The project goal is to support similarity join for billion-scale datasets on a single machine.

- Identify that disk access dominates the execution time of disk-based vector SJ due to read amplification and repetitive data access. Propose DiskJoin, a disk-based vector SJ algorithm optimized for disk scenarios.
- Design smart bucket-wise processing of SJ that reduces disk access drastically, through key contributions of access batching to eliminate read amplification and task orchestration to reduce repetitive data access.
- Introduce a pruning technique that can effectively prune candidate pairs to further accelerate execution.
- **Research Output**: Paper accepted at **SIGMOD 2026**

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**

AWARDS

Champion, SIGMOD'24 Programming Contest — 2024
3rd Place, NeurIPS'23 Big-ANN Competition OOD Track — 2023
Finalist, SIGMOD'23 Programming Contest — 2023

PUBLICATIONS

1. *DiskJoin: Large-scale Vector Similarity Join with SSD*, To appear at **SIGMOD 2026** [Preprint]
   Yanqi Chen, Xiao Yan, Alexandra Meliou, Eric Lo