A Brief Report about the τ-MG Paper

Cong Feng

B.Eng. of BUAA,
M.Eng. candidate of HRBUST

Mar 2023

Outline

The background of ANNS

The main idea of τ-MG

Insights and questions

Approximate Nearest Neighbor Search (ANNS)

Values

- Information retrieval (images/documents/songs).
- Machine learning. e.g., kNN classification and regression.
- Recommendation systems.

Challenges

- Growing size of real-world databases. i.e., from million to billion.
- The curse of dimensionality. i.e., data sparsity in high-dimensional spaces.
- Many tradeoffs. i.e., balancing search/index complexities, time/space complexities, accuracy/latency, etc.

Definition of ANNS

Given a database D of n points in m-dimensional space E^m and a query point $q \in E^m$, the goal of the ANNS is to find a point $p \in D$ s.t. $\delta(q, p) \leq (1 + \epsilon)\delta(q, \bar{v})$, where $\epsilon \geq 0$ is a small constant and $\bar{v}$ is the *exact* nearest neighbor of q.

Taxology of ANNS methods

- Tree-based. i.e., KD-tree, Ball-tree.
- Hashing-based. i.e., Locality Sensitive Hashing (LSH).
- Quantization-based. i.e. Product Quantization (PQ),
- Proximity graph-based i.e., DG, NSWG, RNG, MRNG.

Non-PG methods

- Tree-based methods try to partition (index) the space.
- Hashing-based methods try to retain local similarities in the hamming space.
- They tend to scatter the neighborhood of a node into cells and thus tend to check more points during a search.

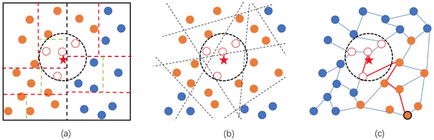


Figure: The index structures and an example search route of (a) tree-based, (b) hashing-based, and (c) PG-based methods. [2]

PG-based methods

- Lacking an error guarantee or search complexity guarantee. i.e., DPG [4], HNSW [5].
- An impractical assumption that $q \in D$. i.e., MRNG [2].
- A high search time complexity due to a long path length. i.e., FANNG [3], MRNG [2], SSG [1].

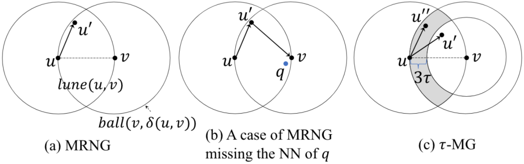


Figure: The edge occlusion rules of (a) MRNG and (c) τ-MG and a failure case of MRNG when $q \notin D$. [6]

Main results

1. Proves the length of any greedy routing path in general PGs is $O(n^{2/m} \ln n)$ with probability at least $1 - (1/e)^{\frac{m}{4}\left(1-\frac{3}{e^2}\right)}$.
2. Proposes τ-MG and a new greedy routing algorithm that guarantees to find $\bar{v}$ in $O(n^{1/m}(\ln n)^2)$ with a great probability given $\delta(q, \bar{v}) < \tau$.
3. Proposes τ-MNG , an approximation of τ-MG with low index complexity and a high chance of finding the exact $\bar{v}$.
4. Proposes three optimizations to aid the performance bottleneck of τ-MNG . (i.e., QEO, PDP, and PII)

Key concepts of τ-MG

τ-monotonic path

A path P is τ-monotonic for a query q if each step of P gets closer to q by at least τ except the last step.

τ-monotonic property

A PG G is τ-monotonic if for any query q given $\delta(q, \bar{v}) < \tau$, G has a τ-monotonic path starting from any point $p \in G$ to $\bar{v}$.

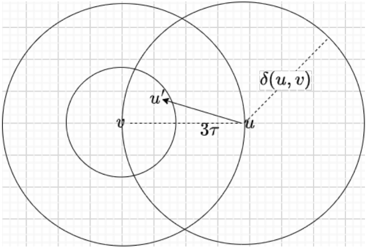
τ-monotonic graph

A τ-MG is τ-monotonic with its edge occlusion rules.

Definition of τ-MG

1. If $\delta(u, v) \leq 3\tau$ then $(u, v) \in G$.
2. Otherwise, either $(u, v) \in G$ or there exists $u' \in ball(u, \delta(u, v)) \cap ball(v, \delta(u, v) - 3\tau)$ s.t. $(u, u') \in G$.

Below is an illustration of the case when $\delta(u, v) > 3\tau$ and the edge (u, u') occludes the edge (u, v).



The idea of τ-MG

1. Existing PG-based methods have a long routing path $O(n^{2/m} \ln n)$ since each step of the path only proceeds $\Delta \leq O((1/n)^{1/m})$. (Theorem 1)
2. A τ-monotonic path has an expected length $O(n^{1/m} \ln n)$ since each step proceeds at least τ (Theorem 2).
3. τ-MG guarantees the existence of a τ-monotonic path from any starting point to the NN of q given $\delta(q, \bar{v}) < \tau$. (Lemma 2)
4. The greedy routing of τ-MG guarantees to find a τ-monotonic path from any starting point to the NN of q. (Lemma 5)

The result

τ-MG guarantees to find the exact NN in time $O(n^{1/m}(\ln n)^2)$ given $\delta(q, \bar{v}) < \tau$.

Theorem 1

The proof first follows the framework in [2]. Then, a small distance $\sqrt{m}\left(\frac{m}{ng}\right)^{1/m}$ is proved to exist with probability at least $1-\left(\frac{1}{e}\right)^{\frac{m}{4}\left(1-\frac{3}{e^2}\right)}$. Thus, Δ is bounded by this value with a great probability.

Theorem 2

The proof is similar to Theorem 1 with the constant τ replacing Δ.

Lemma 2

The proof is done by case analysis.

1. If $(v_0, \bar{v}) \in G$, proof is done trivially.
2. Otherwise, if $\delta(v_0, \bar{v}) < 6\tau$, then v_0 has a neighbor v_1 s.t. $\delta(v_1, \bar{v}) < 3\tau$. Thus, $(v_1, v_0) \in G$ and it can be shown that the path $[v_0, v_1, \bar{v}]$ is τ-monotonic by geometric properties.
3. If $\delta(v_0, \bar{v}) \geq 6\tau$, there exists v_1 s.t. $\delta(v_1, \bar{v}) < \delta(v_0, \bar{v}) - 3\tau$, so it gets closer to $\bar{v}$ by 3τ and closer to q by τ (this is similar to the second case). This continues until for some i it has $\delta(v_i, \bar{v}) < 6\tau$, which falls into the second case.

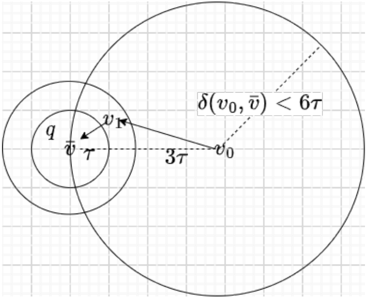


Figure: An illustration of the second case when $\delta(v_0, \bar{v}) < 6\tau$. It can be verified that the path $[v_0, v_1, \bar{v}]$ is τ-monotonic.

Lemma 5

The proof is done by contradiction.

1. If $\bar{v} \notin ball(u, 3\tau)$ then either $(u, \bar{v}) \in G$ or there exists $u' \in ball(u, \delta(u, \bar{v})) \cap ball(\bar{v}, \delta(u, \bar{v}) - 3\tau)$ s.t. $(u, u') \in G$.
2. In the case of $(u, \bar{v}) \in G$, $\bar{v}$ is not farther from q than u, which is a contradiction.
3. In the second case, it can be verified from geometric properties that $\delta(u', q) < \delta(u, q)$, which is also a contradiction.

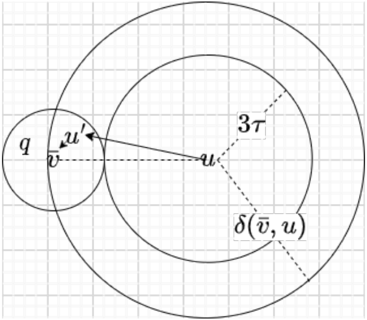


Figure: An illustration of the second case when there exists an edge $(u, u') \in G$ occluding $(u, \bar{v})$. It can be verified that $\delta(u', q) < \delta(u, q)$.

The pros of τ-MG

- Low search latency. It is the fastest PG method so far.
- Practical assumption. $\delta(q, \bar{v}) < \tau$ holds for most real-world databases.

The cons of τ-MG

- Larger index space. It takes $O(n \ln n)$ index space but τ-MNG mitigates this problem [1].
- High index time of $O(n^2 \ln n)$. But τ-MNG lowers it to $O(nh^2 \ln h)$.

[1] The index space complexity of τ-MNG is much lower than τ-MG .

Insights

- It is possible to shorten the expected routing path by a carefully designed PG and a proper assumption.
- The monotonic property of MSNET [2] only guarantees to find the exact NN but not a lower search time. [2]
- A useful technique for the approximation of PGs is the restriction from considering two arbitrary nodes to one node and its arbitrary neighbor.

[2] It is shown that the path length expectation on a general PG is almost the same as that on MSNET.

Q1: Worst-case query

What happens if $\delta(q, \bar{v}) > \tau$ for some q? When a new user enters a system or when there aren't enough points in the database, $\delta(q, \bar{v})$ is likely to be large. What can we do in this case?

Q2: Approximation gaps

How to analyze the performance gaps between τ-MG and its approximation τ-MNG since it is impractical to directly construct τ-MG on large-scale databases?

Q3: Online τ-MG

τ-MG is very suitable for searching in large-scale offline databases (i.e., massive points have been collected in D). Is it meaningful to consider an online version of τ-MG ? (i.e., the updates are very frequent or the usage pattern is *insert if not found*.)

Q4: Larger step of routing

Is it possible that each step of the path proceeds $O(\ln n)$ instead of τ? Since τ can vary for different databases, can it be a function of n?

Thanks for a Patient Hearing!

Reference

[1] Cong Fu, Changxu Wang, and Deng Cai.
High dimensional similarity search with satellite system graph: Efficiency, scalability, and unindexed query compatibility.
IEEE Trans. Pattern Anal. Mach. Intell., 44(8):4139–4150, 2022.

[2] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai.
Fast approximate nearest neighbor search with the navigating spreading-out graph.
Proc. VLDB Endow., 12(5):461–474, 2019.

[3] Ben Harwood and Tom Drummond.
FANNG: Fast Approximate Nearest Neighbour Graphs.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5713–5722, Las Vegas, NV, USA, June 2016. IEEE.

[4] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin.
Approximate nearest neighbor search on high dimensional data - experiments, analyses, and improvement.
IEEE Trans. Knowl. Data Eng., 32(8):1475–1488, 2020.

[5] Yury A. Malkov and Dmitry A. Yashunin.
Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs.
IEEE Trans. Pattern Anal. Mach. Intell., 42(4):824–836, 2020.

[6] Yun Peng, Byron Choi, Tsz Nam Chan, Jianye Yang, and Jianliang Xu.
Efficient Approximate Nearest Neighbor Search in Multi-dimensional Databases.
Proc. the ACM SIGMOD International Conference on Management of Data (SIGMOD '23), June 2023.