Manifold-Consistent Graph Indexing: Overcoming the Euclidean-Geodesic Mismatch via Local Intrinsic Dimensionality

Dongfang Zhao [1]

Abstract

Retrieval-augmented generation (RAG) and approximate nearest neighbor (ANN) search have been critical components of modern large language model (LLM) serving services as they enable efficient and effective retrieval of relevant information from large-scale datasets. However, state-of-the-art methods are mostly based on graph indexing techniques that are agnostic to the intrinsic geometry of the data, and thus often perform poorly in high-dimensional spaces due to a Euclidean-Geodesic mismatch. To that end, we propose a new graph indexing method called Manifold-Consistent Graph Indexing (MCGI). The key idea of MCGI is to leverage the local intrinsic dimensionality of the data to construct a graph that is consistent with the underlying manifold structure, thereby reducing the mismatch and improving performance. Our theoretical analysis shows that MCGI achieves improved approximation guarantees comparing to existing methods, such as HNSW and DiskANN. We also report experimental results demonstrating that MCGI outperforms existing methods in various benchmarks and real-world applications.

1. Introduction

2. Related Work

3. Methodology

3.1. Notations and Definitions

Definition 3.1 (Local Intrinsic Dimensionality)**.** (Houle, 2017) [Dongfang: TODO] Let $\mathcal{X}$ be a domain equipped with a distance measure $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. For a reference point $x \in \mathcal{X}$, let $F_x(r) = \mathbb{P}(d(x, Y) \leq r)$ denote the cumulative distribution function (CDF) of the distance between x and a random variable Y drawn from the underlying data distribution. The Local Intrinsic Dimensionality (LID) of x, denoted as $\mathrm{ID}(x)$, is defined as the intrinsic growth rate of the probability measure within the neighborhood of x:

$$\mathrm{ID}(x) \triangleq \lim_{r \to 0} \frac{r \cdot F'_x(r)}{F_x(r)} = \lim_{r \to 0} \frac{d \ln F_x(r)}{d \ln r} \qquad (1)$$

provided the limit exists and $F_x(r)$ is continuously differentiable for $r > 0$.

4. Evaluation

5. Conclusion

References

Fu, C., Cai, C., Zhou, D., Liu, W., and Wang, C. Fast approximate nearest neighbor search with the navigating spreading-out graph. *Proceedings of the VLDB Endowment*, 12(5):461–474, 2019.

Houle, M. E. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In Beecks, C., Borutta, F., Kröger, P., and Seidl, T. (eds.), *Similarity Search and Applications - 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings*, volume 10609 of *Lecture Notes in Computer Science*, pp. 64–79. Springer, 2017. doi: 10.1007/978-3-319-68474-1_5. URL `https://doi.org/10.1007/978-3-319-68474-1_5`.

Jayaram Subramanya, S., Devvrit, F., Simhadri, H. V., Krishnawamy, R., and Kadekodi, R. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in Neural Information Processing Systems*, 32, 2019.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. In *IEEE transactions on pattern analysis and machine intelligence*, volume 42, pp. 824–836. IEEE, 2018.

[1] University of Washington. Correspondence to: Dongfang Zhao <dzhao@uw.edu>.

Preprint. December 19, 2025.