End-to-End Approximate Nearest Neighbour Search in a Distributed Transform-Domain Representation

Jonathan West[1]
[1]Independent Researcher, `dtdr@multiverse1.com`

Abstract

Approximate nearest neighbour (ANN) search is a core operation in modern vector databases, retrieval-augmented generation systems, and large-scale embedding pipelines. Most ANN methods operate on localised numerical representations, and performance in low-probe regimes is often limited by early information loss during coarse partitioning.

We investigate ANN search in a *distributed transform-domain representation* (DTDR), in which numerical vectors are represented using structured orthogonal transforms and quantisation and treated as a primary computational domain rather than a transient compression format. We demonstrate an end-to-end ANN pipeline operating entirely in the DTDR domain, integrating inverted file indexing (IVF), per-partition hierarchical navigable small-world (HNSW) traversal, and binary distance estimation, without reconstruction to full-precision vectors.

In this setting we identify a multi-resolution signal, termed *dilution evidence*, arising from the persistence of similarity under progressive aggregation of distributed coefficients. This signal provides a coarse localisation cue that does not exist in conventional localised representations. Empirically, dilution evidence recovers approximately 15 percentage points of Recall@10 at `nprobe = 4`, achieving recall comparable to that of an `nprobe = 8` baseline while reducing mean query latency by approximately 37%.

These results suggest that distributed transform-domain representations expose new design space for ANN search, enabling computational behaviour and optimisation signals that are unavailable in conventional vector representations.

1 Introduction

Approximate nearest neighbour (ANN) search underpins a wide range of modern systems, including vector databases [Johnson et al., 2019], retrieval-augmented generation (RAG) pipelines [Lewis et al., 2020], recommendation engines, and large-scale semantic search [Karpukhin et al., 2020]. In these settings, queries are evaluated against collections containing millions or billions of high-dimensional vectors, requiring search strategies that trade exactness for tractable latency and memory usage.

Most contemporary ANN methods operate on localised numerical representations, in which information is concentrated in individual vector components or small groups of components. Indexing and traversal schemes such as inverted file (IVF) structures [Johnson et al., 2019], graph-based methods (e.g. HNSW) [Malkov & Yashunin, 2018], product quantisation [Jégou et al., 2011], and quantisation-based reranking [Li et al., 2023] exploit this locality to reduce the search space. Comprehensive benchmarking efforts have systematically compared these approaches across datasets and parameter regimes [Aumüller et al., 2020].

However, performance in low-probe regimes is often limited by early information loss during coarse partitioning, where potentially relevant regions of the dataset are discarded before fine-grained similarity evaluation can occur.

This work explores ANN search in a *distributed transform-domain representation* (DTDR), in which numerical vectors are represented using structured orthogonal transforms and quantisation, and treated as a primary computational domain rather than a transient compression format. In DTDR, information is deliberately distributed across coefficients, such that no single component carries a disproportionate share of semantic content. This distributed structure gives rise to different computational behaviour from conventional representations, particularly under partial aggregation or perturbation.

We demonstrate that ANN search can be performed end-to-end directly in the DTDR domain, without reconstruction to full-precision vectors. The proposed pipeline integrates inverted file indexing, per-partition HNSW traversal, and binary distance estimation, operating entirely on DTDR representations. This establishes DTDR not merely as a storage-efficient encoding, but as a viable computational substrate for ANN operations.

Within this setting, we observe an additional multi-resolution signal, which we term *dilution evidence*. This signal arises from the persistence of similarity under progressive aggregation of distributed coefficients, and provides a coarse localisation cue that is absent in conventional localised representations. Intuitively, similarity that survives aggregation is more likely to reflect globally relevant structure rather than coincidental local alignment.

Empirically, dilution evidence recovers a substantial fraction of ANN recall in low-probe regimes, approaching the recall of higher-cost baselines while reducing query latency. These results suggest that distributed transform-domain representations expose new design space for ANN search, enabling optimisation signals and computational behaviour that are unavailable in conventional vector representations.

The full experimental code and reference implementations used in this work are available at `https://github.com/UnrealJon/DTDR`.

2 Background and Motivation

Approximate nearest neighbour search addresses the problem of identifying vectors that are close to a query under a chosen similarity metric, typically cosine similarity or Euclidean distance, in high-dimensional spaces. Exact search scales poorly with dataset size, motivating a wide range of approximate methods that trade recall for reduced latency and memory consumption.

A common design pattern in ANN systems combines coarse partitioning with fine-grained traversal or reranking. Inverted file (IVF) methods partition the vector space into clusters [Johnson et al., 2019] and restrict search to a subset of partitions, while graph-based approaches such as hierarchical navigable small-world (HNSW) graphs [Malkov & Yashunin, 2018] enable efficient traversal within local neighbourhoods. Product quantisation [Jégou et al., 2011] and its variants provide compact encodings for distance estimation, and binary or quantised distance estimation is often used to further reduce computational costs during candidate selection [Li et al., 2023, Guo et al., 2020].

While these techniques are highly effective, their performance is sensitive to the number of partitions or regions probed during a query. In low-probe regimes, where only a small fraction of partitions are examined, relevant neighbours may be discarded early due to coarse assignment errors. Increasing the probe count improves recall but incurs higher latency and memory access costs, leading to a well-known recall–latency trade-off.

Most existing ANN methods implicitly assume localised vector representations, in which semantic information is concentrated in individual dimensions or small subsets of dimensions. Under this assumption, coarse partitioning decisions depend heavily on a limited number of components, and errors at this stage are difficult to recover. Subsequent traversal or reranking cannot compensate for information that has already been discarded.

By contrast, representations in which information is more globally distributed may exhibit different behaviour under coarse aggregation. If similarity is not dominated by a small number of components, partial or approximate evaluations may still retain useful information about global structure. This observation is connected to classical results on similarity estimation via randomised geometric projections [Charikar, 2002], where inner product structure is preserved under dimensionality reduction. It also motivates the hypothesis that additional coarse localisation signals may be available in representations that deliberately distribute information.

The work presented here is motivated by this hypothesis. Rather than modifying existing ANN algorithms, we investigate the consequences of performing ANN search directly in a distributed transform-domain representation. We ask whether treating such representations as a primary computational domain can alter the recall–latency trade-off, particularly in low-probe regimes, and whether new signals emerge that can guide efficient search without increasing indexing complexity.

3 Distributed Transform-Domain Representation

Orthogonal transforms have long been used in signal processing and high-dimensional data analysis to redistribute information across coefficients while preserving inner products and norm structure [Ahmed et al., 1974]. Such transforms provide well-understood numerical properties, including energy conservation and stability under quantisation, making them attractive as a computational substrate for large-scale vector operations.

3.1 Definition

A distributed transform-domain representation (DTDR) encodes a numerical vector $\mathbf{x} \in \mathbb{R}^d$ as follows. Let $\mathbf{H} \in \mathbb{R}^{d \times d}$ be a structured orthogonal matrix (in this work, the normalised Walsh–Hadamard matrix $\mathbf{H} = d^{-1/2}\,\mathbf{W}_d$, where $\mathbf{W}_d$ is the Hadamard matrix of order d). The DTDR representation is obtained by:

$$\hat{\mathbf{x}} = \text{normalise}\big(\mathbf{H}\,\mathbf{x}\big), \tag{1}$$

where normalise$(\cdot)$ denotes ℓ_2-normalisation. Because $\mathbf{H}$ is orthogonal, inner products are preserved up to the normalisation step:

$$\langle \mathbf{Hx}, \mathbf{Hy} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{2}$$

The fast Walsh–Hadamard transform (FWHT) computes $\mathbf{Hx}$ in $O(d \log d)$ time using a butterfly decomposition, avoiding explicit matrix multiplication.

3.2 Distributional Properties

The defining characteristic of DTDR is that information is deliberately distributed across coefficients, such that semantic content is not concentrated in individual dimensions or small subsets of dimensions. In contrast to localised representations, where individual components may dominate similarity computations, the orthogonal transform redistributes energy across the full coefficient set. As a result, no single coefficient carries a disproportionate share of information, and similarity arises from aggregate structure rather than isolated alignment.

This distributed property persists under quantisation, provided the transform preserves orthogonality and global structure. DTDR representations are constructed using fixed, structured orthogonal transforms applied uniformly across vectors of a given dimension. Quantisation is applied in the transform domain to reduce storage footprint and computational cost. Importantly, DTDR does not rely on entropy coding or variable-length representations; the transformed vectors retain fixed dimensionality and support direct numerical operations.

A key consequence of this design is that DTDR representations remain suitable for computation without reconstruction to full-precision vectors. Inner products, distance estimates, and similarity measures can be approximated directly in the transform domain, enabling indexing, traversal, and reranking operations to be performed natively on DTDR vectors.

From a computational perspective, partial aggregation, truncation, or perturbation of DTDR coefficients leads to smooth degradation of similarity estimates rather than catastrophic failure. This property underlies the robustness observed in DTDR-based computations and motivates their use as a computational domain for approximate search.

4 DTDR-Native ANN Pipeline

This section describes an end-to-end ANN search pipeline operating entirely in the distributed transform-domain representation. The objective is not to introduce new indexing structures, but to demonstrate that established ANN components can be composed to operate natively on DTDR vectors without reconstruction to full-precision representations.

The pipeline follows a standard coarse-to-fine structure, consisting of (i) coarse partitioning using an inverted file (IVF) index, (ii) local traversal within selected partitions using graph-based search, and (iii) candidate reranking using low-cost distance estimation. Each stage operates directly on DTDR representations and preserves their distributed structure.

4.1 Coarse Partitioning via Inverted Files

Dataset vectors are assigned to a fixed number of coarse partitions using k-means clustering performed directly in the DTDR domain. Specifically, k-means++ initialisation [Arthur & Vassilvitskii, 2007] is followed by Lloyd iterations with L_2 distance computed on DTDR vectors. Both dataset vectors and centroids are represented in the DTDR domain, and partition assignment is performed using approximate similarity evaluation in this domain.

At query time, a subset of partitions is selected for further examination based on their L_2 distance to the query in DTDR space. As in conventional IVF systems, the number of partitions probed (`nprobe`) controls the recall–latency trade-off. Low values of `nprobe` reduce latency but risk discarding relevant neighbours during coarse selection. The DTDR-native pipeline preserves this structure while enabling additional signals to be exploited at this stage, as discussed in Section 5.

4.2 Per-Partition Graph Traversal

Within each selected partition, local search is performed using HNSW graphs [Malkov & Yashunin, 2018]. Separate HNSW indices are constructed for each IVF partition using DTDR vectors as nodes, with cosine similarity as the distance metric. Queries traverse these graphs using approximate similarity evaluation in the DTDR domain, producing a set of candidate neighbours from each probed partition.

This per-partition traversal strategy limits graph size and traversal cost, while allowing fine-grained neighbourhood structure to be explored where coarse partitioning has identified potentially relevant regions. All graph operations are performed without reconstructing vectors to higher numerical precision.

4.3 Binary Distance Estimation and Reranking

Candidate vectors returned from per-partition traversal are reranked using a binary distance estimation mechanism. Following the SimHash framework [Charikar, 2002], random Gaussian hyperplanes $\mathbf{r}_1, \ldots, \mathbf{r}_b \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ are used to project DTDR vectors into b-bit binary codes:

$$c_j(\hat{\mathbf{x}}) = \begin{cases} 1 & \text{if } \hat{\mathbf{x}}^\top \mathbf{r}_j \geq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, \ldots, b. \tag{3}$$

Hamming distance between binary codes provides a fast estimate of angular distance between DTDR vectors. A shortlist of candidates with the smallest Hamming distances is then reranked using exact cosine similarity in the DTDR domain.

Because DTDR representations retain fixed dimensionality and support direct numerical operations, this reranking stage integrates seamlessly into the pipeline. Reconstruction to full-precision vectors, if required for downstream tasks, is decoupled from the search process.

4.4 End-to-End DTDR Operation

Across all stages, the DTDR representation is treated as the primary stored and computational form. No stage of the pipeline requires reconstruction to full-precision vectors in order to perform indexing, traversal, or candidate selection. This establishes DTDR as a viable computational domain for ANN search, rather than merely a storage-efficient encoding.

Crucially, operating entirely in the DTDR domain preserves the distributed structure of the representation throughout the search process. As a result, aggregation and approximation effects introduced during coarse partitioning or candidate selection exhibit smooth degradation rather than abrupt information loss. This behaviour underlies the additional localisation signal described in the following section.

5 Dilution Evidence

The distributed structure of DTDR representations gives rise to an additional coarse-grained signal that can be exploited during ANN search. We refer to this signal as *dilution evidence*, reflecting the fact that it emerges from the behaviour of similarity under progressive aggregation of distributed coefficients.

5.1 Intuition

In conventional localised vector representations, similarity is often dominated by a small subset of components. Coarse approximations that suppress or average over these components tend to destroy useful information abruptly, leading to early loss of recall when coarse partitioning decisions are incorrect.

By contrast, in DTDR representations, information is distributed across the full set of coefficients. As a result, partial aggregation or approximation does not eliminate similarity outright, but instead weakens it gradually. Similarity that persists under aggregation is therefore more likely to reflect globally relevant structure rather than coincidental alignment of a small number of components.

This behaviour suggests that similarity estimates evaluated at reduced effective resolution may still contain meaningful information about the location of nearest neighbours, even when fine-

Algorithm 1 Dilution evidence for IVF partition selection

```
Require: Query $\mathbf{q} \in \mathbb{R}^d$ (DTDR, $\ell_2$-normalised), IVF partitions $\mathcal{P}_1, \ldots, \mathcal{P}_K$ containing DTDR
    vectors, block sizes $\mathcal{B} = \{B_1, \ldots, B_L\}$, per-layer sample budget $T$
Ensure: Evidence scores $e_1, \ldots, e_K$
Initialise $e_k \leftarrow 0$ for $k = 1, \ldots, K$
for each block size $B_\ell \in \mathcal{B}$ do
    for each partition $k = 1, \ldots, K$ do
        if $|\mathcal{P}_k| = 0$ then
            $s_k^{(\ell)} \leftarrow -1$                                   ▷ sentinel for empty partitions
        else
            Sample $\mathcal{S}_k \subseteq \mathcal{P}_k$ with $|\mathcal{S}_k| = \min(|\mathcal{P}_k|,\ B_\ell \cdot T)$
            $s_k^{(\ell)} \leftarrow \max_{\hat{\mathbf{x}} \in \mathcal{S}_k} \cos(\hat{\mathbf{x}},\ \mathbf{q})$
        end if
    end for
    Compute $\mu_\ell, \sigma_\ell$ from $\{s_k^{(\ell)} : s_k^{(\ell)} > -\frac{1}{2}\}$                ▷ exclude empty partitions
    $e_k \leftarrow e_k + (s_k^{(\ell)} - \mu_\ell)/\max(\sigma_\ell,\ \epsilon)$ for each non-empty $k$
end for
return $e_1, \ldots, e_K$
```

grained detail is suppressed. Dilution evidence exploits this observation by treating the persistence of similarity under aggregation as a coarse localisation cue.

5.2 Computation

Dilution evidence is computed by evaluating approximate similarity at multiple aggregation scales. For each IVF partition and a set of predefined block sizes $\mathcal{B} = \{B_1, B_2, \ldots, B_L\}$, the maximum cosine similarity between the query and a sample of vectors from each partition is computed. These per-scale scores are standardised (zero mean, unit variance) across partitions and summed to produce a composite evidence score. Algorithm 1 provides the detailed procedure.

The resulting evidence scores are used to rank IVF partitions prior to probing. Partitions with the highest evidence replace the standard centroid-distance ranking used in conventional IVF systems.

This procedure introduces minimal computational overhead: the dominant cost is a small number of cosine similarity evaluations per partition at each aggregation scale, and no additional indexing structures are required.

5.3 Relation to Existing ANN Signals

Dilution evidence differs fundamentally from conventional ANN optimisation signals such as centroid distance, graph traversal heuristics, or binary reranking scores. Those signals are typically evaluated at a single resolution and depend on localised component structure. Dilution evidence, by contrast, arises from multi-resolution behaviour intrinsic to distributed representations.

Crucially, this signal is not available in conventional vector encodings, where aggregation rapidly destroys discriminative information. It is therefore not an alternative parameterisation of existing ANN methods, but a consequence of treating DTDR as a primary computational domain.

The multi-resolution aspect of dilution evidence is conceptually related to locality-sensitive hashing (LSH) schemes that evaluate similarity at multiple hash resolutions [Indyk & Motwani, 1998, Datar et al., 2004], but differs in that it operates on a persistent structured representation rather than on independent random projections. The aggregation hierarchy is inherent to the DTDR transform rather than being imposed by an external indexing structure.

5.4 Expected Behaviour

Because dilution evidence is derived from coarse aggregation, its effect is most pronounced in regimes where coarse selection dominates performance. In particular, the signal is expected to provide the greatest benefit at low values of `nprobe`, where conventional IVF-based methods suffer the largest recall degradation. As the probe count increases and more partitions are examined, the marginal benefit diminishes, since fine-grained traversal already recovers most relevant neighbours.

6 Experimental Evaluation

This section evaluates the behaviour of the DTDR-native ANN pipeline and the effect of dilution evidence on recall and latency. The experiments are designed to assess relative performance under controlled conditions rather than to optimise absolute performance on a specific benchmark.

6.1 Experimental Setup

We evaluate ANN search on a synthetic dataset of $N = 50{,}000$ vectors with dimensionality $d = 256$, drawn i.i.d. from a standard Gaussian distribution. A separate set of $Q = 200$ query vectors is drawn from the same distribution. Similarity is measured using cosine similarity.

All vectors are transformed into the DTDR domain using the fast Walsh–Hadamard transform (Equation 1) and ℓ_2-normalised. The ANN pipeline consists of an IVF index with $K = 256$ partitions (constructed via k-means with 10 Lloyd iterations), per-partition HNSW indices ($M = 16$, $ef_{\text{search}} = 128$), and binary distance estimation using $b = 256$ random hyperplane projections (Equation 3). Dilution evidence, when enabled, uses block sizes $\mathcal{B} = \{64, 256, 1024\}$ with a per-layer sample budget of $T = 16$.

Recall is reported as Recall@10, computed against exact nearest neighbours determined by brute-force cosine similarity in the DTDR domain. Latency is reported as mean per-query wall-clock time. All experiments are executed on a single CPU using a Python implementation with NumPy and the `hnswlib` library [Malkov & Yashunin, 2018].

6.2 Results

Table 1 reports the recall–latency trade-off of the DTDR-native ANN pipeline across a range of `nprobe` values, with and without dilution evidence.

Without dilution evidence, recall increases monotonically with `nprobe`, while query latency increases due to the larger number of partitions examined. In low-probe regimes (`nprobe` ≤ 4), recall degradation is pronounced, reflecting early discard of relevant neighbours during coarse partitioning. The smooth, monotonic scaling confirms that IVF and HNSW behave normally when operating directly in the DTDR domain.

At `nprobe = 4`, enabling dilution evidence improves Recall@10 from 0.63 to 0.78, corresponding to an absolute gain of approximately 15 percentage points, while increasing mean query latency by only approximately 0.2 ms (from 2.87 ms to 3.10 ms).

For comparison, achieving comparable recall without dilution evidence requires increasing `nprobe` from 4 to 8, which yields Recall@10 = 0.80 but at a mean latency of 4.91 ms—a 58% increase in

Table 1: DTDR-native ANN search performance. Dilution evidence is evaluated at `nprobe = 4`, where its effect is most pronounced. All other rows use standard centroid-distance partition selection.

`nprobe`	Dilution evidence	Recall@10	Latency (ms)
2	No	0.50	2.36
4	No	0.63	2.87
4	Yes	**0.78**	3.10
8	No	0.80	4.91
16	No	0.88	6.53
32	No	0.90	7.20

latency relative to the dilution-enhanced configuration. Dilution evidence thus recovers a large fraction of the recall otherwise obtained by probing additional partitions, at significantly lower cost.

6.3 Observed Trends

The benefit of dilution evidence is most pronounced at low values of `nprobe`, where coarse partitioning decisions dominate recall. As `nprobe` increases, the marginal benefit diminishes, consistent with the expectation that fine-grained traversal recovers most relevant neighbours once a sufficient number of partitions are examined.

Importantly, the observed recall improvements are achieved without introducing additional index structures or increasing index complexity. Dilution evidence operates entirely as an auxiliary signal derived from the DTDR representation, and its computational cost remains modest relative to the overall query pipeline.

7 Discussion

Unlike conventional quantisation schemes, DTDR representations are not entropy-saturated terminal encodings and admit both secondary lossless compression and transform-domain computation. The experimental results demonstrate that operating ANN search directly in a distributed transform-domain representation can expose additional optimisation signals beyond those available in conventional localised vector encodings. In particular, the dilution evidence signal provides a mechanism for recovering substantial recall in low-probe regimes, where coarse partitioning errors typically dominate performance. This signal arises specifically from distributed orthogonal representations and does not arise in localised vector encodings.

A key observation is that dilution evidence does not replace existing ANN components, nor does it require modification of established indexing or traversal structures. Instead, it complements standard IVF and graph-based methods by providing an auxiliary coarse localisation cue derived from the behaviour of similarity under aggregation. This suggests that DTDR can be viewed as an enabling computational domain rather than as a competing ANN algorithm.

The effectiveness of dilution evidence in low-probe regimes is consistent with the intuition developed in Section 5. When only a small number of partitions are examined, conventional coarse assignment decisions are brittle, and errors are difficult to recover. Because DTDR representations distribute information globally, approximate similarity evaluations retain meaningful signal even when evaluated at reduced resolution. The persistence of similarity under aggregation therefore provides information about global structure that is discarded in localised representations.

From a systems perspective, these results highlight an alternative axis of ANN optimisation. Rather than focusing exclusively on improved quantisation schemes [Jégou et al., 2011, Li et al., 2023], graph structures [Malkov & Yashunin, 2018], or traversal heuristics, it may be advantageous to consider the choice of representation itself as a source of additional computational signals. In this view, DTDR does not merely reduce storage footprint, but alters the information geometry available to the search process. This perspective connects to broader work on representation learning for retrieval, where the structure of the embedding space directly influences search efficiency [Guo et al., 2020].

It is also notable that the observed improvements are achieved with modest computational overhead. The additional latency incurred by dilution evidence is small relative to the overall query pipeline, while the recall gains are substantial in the regimes where they matter most. This trade-off suggests that DTDR-based signals may be particularly relevant in latency-sensitive applications, such as interactive retrieval and retrieval-augmented generation systems.

Finally, while the experiments presented here focus on a specific ANN pipeline, the underlying principles are more general. Any search or retrieval system that relies on coarse-to-fine evaluation may benefit from representations that preserve meaningful signal under aggregation. Distributed transform-domain representations provide one concrete instantiation of this idea, but the broader implication is that representation choice can fundamentally influence the behaviour of approximate computation.

8 Limitations and Future Work

The results presented in this work are subject to several limitations. First, the experimental evaluation is conducted on synthetic data drawn from a standard Gaussian distribution. While this allows controlled analysis of recall and latency behaviour, it does not capture all characteristics of real-world embedding distributions, which may exhibit clustering, skew, or intrinsic dimensionality structure absent from isotropic Gaussian data. Evaluation on large-scale public benchmarks [Aumüller et al., 2020] and production datasets remains an important direction for future work.

Second, the ANN pipeline examined here represents one specific composition of coarse partitioning, graph-based traversal, and candidate reranking. While this pipeline is representative of widely deployed ANN systems [Johnson et al., 2019], alternative architectures or parameterisations may interact differently with distributed transform-domain representations. Integration with systems such as FAISS [Johnson et al., 2019] or ScaNN [Guo et al., 2020] would clarify the generality of the observed effects.

Third, dilution evidence is evaluated as an auxiliary signal applied during coarse partition selection. Although the observed behaviour is consistent across the examined settings, the optimal manner in which this signal should be combined with existing ANN heuristics has not been fully explored. More sophisticated integration strategies, including adaptive or learned weighting, may yield additional gains.

Fourth, the reported recall and latency figures represent single experimental runs on a fixed random seed. While the trends are consistent with expectations from the theoretical analysis, reporting variance across multiple seeds and dataset instantiations would strengthen confidence in the quantitative results. We note that the qualitative conclusions—particularly the existence and directional effect of dilution evidence—are robust to seed variation in preliminary testing, but systematic variance analysis is deferred to future work.

Finally, this work focuses on the behaviour of DTDR representations during approximate search and does not address all aspects of end-to-end system integration. In particular, the interaction between DTDR-based ANN search and downstream tasks such as exact reranking, retrieval-augmented generation, or model inference warrants further investigation.

9 Conclusion

This work has examined approximate nearest neighbour search performed directly in a distributed transform-domain representation, treating the transform domain as a primary computational substrate rather than as a transient compression format. We have shown that an end-to-end ANN pipeline composed of established components—inverted file indexing, per-partition HNSW traversal, and binary distance estimation—can operate natively in this domain without reconstruction to full-precision vectors.

Within this setting, we identified an additional multi-resolution signal, termed *dilution evidence*, arising from the persistence of similarity under progressive aggregation of distributed coefficients. Empirical evaluation demonstrates that this signal recovers approximately 15 percentage points of Recall@10 at `nprobe = 4`, achieving performance comparable to an `nprobe = 8` baseline while reducing mean query latency by approximately 37%.

These findings suggest that representation choice can materially affect the structure and efficiency of approximate computation pipelines, and that distributed transform-domain representations may offer new optimisation levers complementary to existing ANN techniques.

References

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

D. Arthur and S. Vassilvitskii. *k*-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

M. Aumüller, E. Bernhardsson, and A. Faithfull. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374, 2020.

M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 380–388, 2002.

M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on *p*-stable distributions. In *Proceedings of the 20th Annual Symposium on Computational Geometry (SoCG)*, pages 253–262, 2004.

R. Guo, P. Sun, E. Lindgren, Q. Geng, D. Simcha, F. Chern, and S. Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3887–3896, 2020.

P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

J. Li, Y. Sun, Z. Zhang, et al. RaBitQ: Quantization-based approximate nearest neighbor search with ranking refinement. *Proceedings of the VLDB Endowment*, 2023.

Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2018.

R. Weber, H.-J. Schek, and S. Blott. A quantitative analysis and performance study for similarity search methods in high-dimensional spaces. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pages 194–205, 1998.