

AiSAQ : All-in-Storage ANNS with Product Quantization for DRAM-free Information Retrieval

Kento Tatsuno*

kento1.tatsuno@kioxia.com
Kioxia Corporation
Japan

Kiyoshi Ishiyama

kiyoshi1.ishiyama@kioxia.com
Kioxia Corporation
Japan

Daisuke Miyashita*

daisuke1.miyashita@kioxia.com
Kioxia Corporation
Japan

Taiga Ikeda

taiga1.ikeda@kioxia.com
Kioxia Corporation
Japan

Jun Deguchi

jun.deguchi@kioxia.com
Kioxia Corporation
Japan

ABSTRACT

Graph-based approximate nearest neighbor search (ANNS) algorithms work effectively against large-scale vector retrieval. Among such methods, DiskANN achieves good recall-speed tradeoffs using both DRAM and storage. DiskANN adopts product quantization (PQ) to reduce memory usage, which is still proportional to the scale of datasets. In this paper, we propose All-in-Storage ANNS with Product Quantization (AiSAQ), which offloads compressed vectors to the SSD index. Our method achieves ~10 MB memory usage in query search with billion-scale datasets without critical latency degradation. AiSAQ also reduces the index load time for query search preparation, which enables fast switch between multiple billion-scale indices. This method can be applied to retrievers of retrieval-augmented generation (RAG) and be scaled out with multiple-server systems for emerging datasets. Our DiskANN-based implementation¹ is available on GitHub.

1 INTRODUCTION

In tasks of retrieving specific data that users are interested in from large-scale datasets such as images, music, and documents, Approximate Nearest Neighbor Search (ANNS), which gets close vectors to query vectors using pre-built indices constructed from the datasets, is commonly utilized. Usually, the index data is stored in fast memories (RAM) to shorten the search latency. However, the DRAM costs too much to store large datasets, especially those that reach a billion-scale, and slower but cheaper storages such as SSDs also need to be employed.

In the recent trend of Large Language Models (LLM), possible implementations of Retrieval-Augmented Generation (RAG)[10] use vector datasets as their external knowledge sources to generate more accurate answers. In the original paper, Dense Passage Retrieval[9] is used as a “retriever” for searching the external sources. That retriever uses HNSW[11] for searching vector data of external sources, and it can be replaced with other suitable ANNS algorithms depending on the demand of applications. In some situations that a RAG chain requires multiple sources, the retriever need to “switch” between indices. This procedure requires the index data of

all sources to be retained in RAM during the service time or loaded from storage every time of a request.

DiskANN[7] is the de-facto standard of graph-based ANNS methods exploiting storages. This is adopted for the baseline of Big-ANN competition track of NeurIPS[14] and is employed in vector database services such as Weaviate[5] and Zilliz [17]. DiskANN claims to achieve query search of billion-scale datasets with high recall and small memory usage. It reduces memory usage by compressing node vectors using Product Quantization(PQ) [8] and keeps high search recall by re-ranking all nodes in the search path by their full-precision vectors loaded from the storage. However, since DiskANN retains PQ-compressed vectors of all nodes in RAM, the memory usage is proportional to the scale of the datasets. Additionally, increasing compression ratio reduces both of memory usage and recall, which leads to the trade-off between them. This means DiskANN is not truly scalable to the scale of datasets and is not suitable for switching between multiple indices because loading all PQ vectors onto DRAM can take too long time.

In this paper, we propose AiSAQ : All-in-Storage ANNS with Product Quantization , a novel method of index data placement. Our implementation achieved only ~10 MB of DRAM usage regardless of the scale of datasets. Our optimization for data placement in the storage index brings minor latency degradation compared to DiskANN, despite larger portion of the index are sitting in the storage. Moreover, index switch between multiple billion-scale datasets can be performed in millisecond-order and makes cost advantage in multiple-server environments.

The main contributions of this paper are follows:

- AiSAQ query search requires only ~10 MB of RAM for any size of dataset including SIFT1B billion-scale dataset, without changing graph topology and high recall of original DiskANN.
- Thanks to data placement optimization on SSD, AiSAQ achieves millisecond-order latency for >95% 1-recall@1 with tiny memory usage.
- AiSAQ achieved negligible index load time before query search. Applications can flexibly switch between multiple billion-scale corpus corresponding to users’ requests.
- Index switch between different datasets in the same vector space can further reduce the switching time to sub-millisecond order, due to sharing their PQ centroid vectors.

*Both authors contributed equally to the paper.

¹<https://github.com/KioxiaAmerica/aisaq-diskann>



- Multiple-server search for large-scale indices makes cost advantage of total DRAM and SSD consumption.

As a concurrent work, LM-DiskANN[12] was published before this paper’s submission. However, our work proves the scalability mentioning billion-scale datasets and multiple-server systems for emerging size of datasets in the future. We also propose a new metric called index switching time, whose result reinforces the scalability for larger datasets. Both of these points are not stated in their paper.

2 PRELIMINARIES

2.1 Graph-Based ANNS Algorithms

In a dataset X containing N points of d -dimensional vector $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$ and a given query vector $q \in \mathbb{R}^d$, nearest neighbor search (NNS) is a task to compute the nearest neighbor vector $\hat{x} = \operatorname{argmin}_{x_i} d(x, q)$ in a distance metric $d(\cdot, \cdot)$. Being a brute-force computation, NNS takes $O(Nd)$ time complexity and is not optimal for large-scale datasets.

To reduce the query search time, methods to approximately compute nearest neighbor vectors are proposed, which are called approximate nearest neighbor search (ANNS).

Among various approaches of ANNS, a graph-based method uses directed graph as an index. In the index construction phase, the algorithm regards the dataset vectors x_i as a node v_i and makes some edges between nodes based on their distances. In the query search phase, the algorithm starts from the entry point and move along the edges, which navigates to the candidate point of the nearest neighbor \hat{x} of the query q . These graph-based ANNS methods empirically achieve $O(d \log N)$ time complexity[3]. Some ANNS algorithms use compressed vectors like PQ ones instead of full-precision vectors to reduce memory usage and computational cost by compromising precise distance calculation.

2.2 Index Switch for Multiple Sources

Some RAG systems retrieve information from external sources using ANNS methods. For further improvement of LLMs with RAG, ANNS methods need to “switch” between multiple indices of different domains corresponding to users’ or applications’ requests. For example, an LLM which generates latest news across multiple fields such as economy, politics and entertainment needs to generate a single answer with multiple sources of those fields. Another example is an LLM chain which uses multiple sources. Represented by LangChain[4], it is common to use a pipeline with multiple-time sequential LLM inferences which uses the output from an LLM inference as the input for the next inference. In such situations, it is possible that a RAG inference requires a different source than the one used in the preceding inference. Existing ANNS methods have to retain all of the indices on RAM or load data of an index every time from storage to switch between multiple indices. This leads to large DRAM usage or long index load time.

2.3 DiskANN

DiskANN[7], one of graph-based ANNS methods, constructs a flat directed graph index by an algorithm called Vamana. In the query

Algorithm 1 DiskANN Beam Search with re-ranking

Data: Graph G with entrypoint s , query q , number of top candidates k , beamwidth w , and search list size $L \geq k$

Result: PQ compressed node set \mathcal{L} and full-precision node set \mathcal{V} , each containing top- k neighbor candidates of q

```

 $\mathcal{L} \leftarrow \{s\}, \mathcal{V} \leftarrow \emptyset$ 
while  $\mathcal{L} \setminus \mathcal{V} \neq \emptyset$  do
    let  $\mathcal{P}$  denote top- $w$  closest nodes to  $q$  in  $\mathcal{L}$ 
    for  $p \in \mathcal{P}$  do
        get node chunk of  $p$  from storage
    end for
    append  $PQ$  vectors of  $N_{out}(p)$  from RAM to  $\mathcal{L}$ 
    append full-precision vectors of  $p$  from node chunk to  $\mathcal{V}$ 
     $\mathcal{L} \leftarrow \mathcal{L} \cup N_{out}(p), \mathcal{V} \leftarrow \mathcal{V} \cup p$ 
    if  $|\mathcal{L}| > L$  then
        set  $\mathcal{L}$  with top- $L$  closest nodes in  $PQ$  space to  $q$  in  $\mathcal{L}$ 
    end if
end while
sort  $\mathcal{V}$  by their full-precision distance to  $q$ 
return top- $k$  closest nodes to  $q$  in  $\mathcal{V}$ 

```

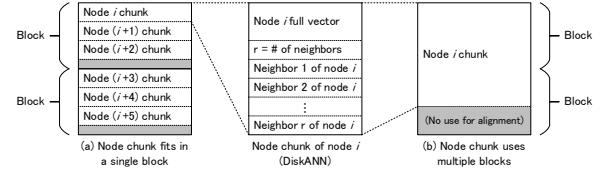


Figure 1: Node chunk details and alignment in LBA blocks

search phase, we traverse the graph according to a search algorithm called beam search derived from greedy search.

Algorithm 1 shows detailed procedure of beam search used in DiskANN. The PQ compressed vector is stored on DRAM and is used for search path determination. On the other hand, the full-precision vector and IDs of outneighbors $N_{out}(v_i)$ of v_i is written in the continuous LBA (Logical Block Address) space on the storage like SSD (Figure 1). In this paper, we call this information chunk of a specific node v_i “node chunk” of v_i . In each graph hop, DiskANN reads the node chunk of v_i and determines the next search path comparing PQ distance between query and each of $N_{out}(v_i)$, looking up PQ vectors retained on the DRAM. After the graph hops finished, DiskANN sorts the candidate nodes in the search path by their full-precision distances, which is called re-ranking.

With b_{full} bytes of each full-precision vector, max outdegree R , and b_{num} bytes (usually 4 bytes) to express a node ID or the outdegree, the size of a single node chunk is $B_{DiskANN} = b_{full} + b_{num}(R + 1)$. Operation systems dispatches I/O requests to storage devices in “blocks”, whose size is $B = 4$ KB in most OS settings. In most cases, $B_{DiskANN} \leq B$ stands and a block contains a single or multiple node chunk(s) in this case (Figure 1 (a)). If $B_{DiskANN} > B$,

the chunk uses multiple blocks (Figure 1 (b)). In both cases, node chunk which doesn't fit in the rest space after previous node chunk in a block is aligned to the start of the next block. The query search of DiskANN dispatches I/O requests to read $\lceil \frac{B_{DiskANN}}{B} \rceil$ block(s) used by a single node chunk. For example, assuming that block size is 4 KB, the I/O size is 4 KB in Figure 1 (a) and 8 KB in Figure 1 (b).

2.4 Drawbacks of DiskANN

Memory Usage. DiskANN loads PQ vectors of all nodes in its dataset. This means memory usage of DiskANN is roughly proportional to the scale of the dataset N . For instance, an index constructed from SIFT1B dataset with 32 byte per PQ vector, which corresponds to 1/4 compression of the original vector, requires 32 GB of RAM. Increasing the PQ compression ratio brings recall degradation even with re-ranking, and the memory usage is still proportional to N .

Spatial Efficiency. In the DiskANN beam search for a query, the number of nodes used for distance computation is much smaller than the dataset scale N . Even for multiple queries, few nodes are accessed frequently, while most nodes are unused and just occupy a large amount of RAM.

Index Switch Time. Conventional graph-based ANNS methods including DiskANN cost time to load vector data to DRAM before the query search starts, which also increases together with N . To realize index switch, using multiple large-scale datasets and switching the corpus corresponding to applications' requests, conventional methods need to keep all vector data of multiple datasets in DRAM to avoid the loading time.

3 PROPOSED METHOD

To address the above problems, we propose a novel method named AiSAQ : All-in-Storage ANNS with Product Quantization which offloads PQ vector data from DRAM to storage and aims almost zero memory usage with negligible latency degradation.

3.1 Methodology

When the DiskANN beam search is located in a node $v(i)$, the PQ vectors of $N_{out}(v_i)$ correspond to their IDs stored in the node chunk of $v(i)$. This means these PQ vectors themselves can also be stored in the node chunk, which is the main idea of AiSAQ (Figure 2). AiSAQ also needs to adjust the node chunks which has larger size than DiskANN to fill blocks effectively and to minimize the I/O latency increment. With b_{PQ} bytes of each PQ vector, the node chunk size of AiSAQ is $B_{AiSAQ} = B_{DiskANN} + Rb_{PQ} = b_{full} + b_{num} + R(b_{num} + b_{PQ})$. It is recommended to adjust the maximum degree R so that $B_{AiSAQ} \leq nB$ or $B_{AiSAQ} \leq \frac{B}{n}$ ($n \in \mathbb{N}$) stands.

In the query search phase, PQ vectors of $N_{out}(v_i)$ are obtained from the node chunk of v_i in the storage instead of memory and used to compute PQ distances to the query. Utilizing unused area of blocks in DiskANN by filling with PQ vectors, this needs no additional I/O requests for getting PQ vectors and enables distance computations of each hop within the node chunk without DRAM. The PQ vectors can be discarded after distance computations in each hop. Since the PQ vector of a specific node is obtained in the previous hop, the above algorithm can't get PQ vectors of the

Table 1: Datasets and index build parameters used for AiSAQ experiments

Dataset	SIFT1M	SIFT1B	KILT E5 22M
# of vectors N	1,000,000	1,000,000,000	22,220,792
Dimensionality d	128	128	1024
Data type	float	uint8	float
Distance metric	Euclid	Euclid	MIPS
Maximum outdegree R	56	52	69
# of PQ subvectors b_{PQ}	128	32	128

entrypoints. Therefore, we keep only PQ vectors of n_{ep} (1 in most cases) entrypoints in DRAM. In this method, we have to keep at most $(R + n_{ep})$ PQ vectors in DRAM at once, which is independent of and much smaller than N , and can aim near-zero memory usage.

3.2 Implementation

We implemented AiSAQ index creation algorithm based on the existing DiskANN programs [15] and its query search algorithm from scratch, only reusing some utilities from DiskANN. From Vamana graph and PQ vectors generated by DiskANN, a single AiSAQ index file is generated. Our implementation prioritized to lower the memory footprint, which does not support additional features of DiskANN such as filtering[6] and dynamic indexing[16].

4 EVALUATION

In this section, we conducted some query search experiments using AiSAQ index files and compared with DiskANN results in the same construct and search conditions.

4.1 Datasets and Experimental Conditions

We benchmarked AiSAQ and DiskANN with various scales of SIFT datasets[2] and KILT E5 dataset, the latter of which is a 1024-dimensional vector dataset generated by applying e5-large-v2[18] embedding to KILT[13] knowledge source.

Table 1 shows specifications of the above datasets used for DiskANN and AiSAQ indices. Each PQ subvector of all experiments in this section is categorized into 256 clusters and can be represented in 8 bits (1 byte), which means the number of PQ subvectors can be translated into the size b_{PQ} in bytes of each PQ compressed vectors. We built one-shot indices, adjusting the maximum degree R for each dataset so that node chunk effectively fills a single block or multiple blocks. In each search, we fixed the beamwidth $w = 4$ and changed the candidate list size L .

We conducted single-server query search experiments (4.2 to 4.4) on an AWS EC2 i4i.8xlarge instance which has 2x 3750 GB instance stores tied up by RAID 0. The environment of multiple-server experiments is detailed in 4.5.

4.2 Memory Usage

We measured peak memory usage of DiskANN and AiSAQ query search by /usr/bin/time command. Since both of the algorithms load the entire query files onto DRAM and that portion is not essential for algorithm comparison itself, measurement of memory

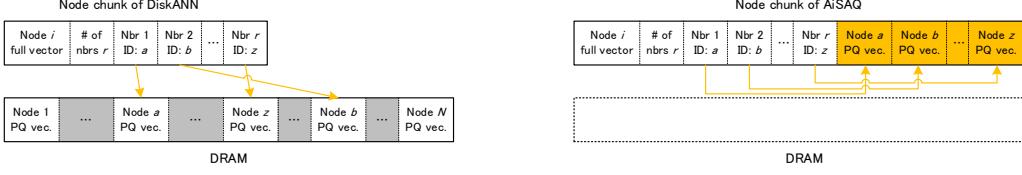


Figure 2: Data placements of a node chunk and memory of DiskANN (left) and proposed method AiSAQ (right)

Table 2: Memory Usage (MB) by DiskANN and AiSAQ

	b_{PQ}	DiskANN	AiSAQ (ours)
SIFT1M	128 bytes	146	11
SIFT1B	32 bytes	31,303	11
KILT E5 22M	128 bytes	2,803	14

usage was executed only for 10 queries without groundtruth data in each dataset. Table 2 shows the memory usage of DiskANN and AiSAQ given the same index construction and search parameters for each dataset. While DiskANN required up to 32 GB of RAM in SIFT1B dataset, AiSAQ used at most 14 MB of RAM even for billion-scale query search.

The breakdown of AiSAQ’s memory usage inspected with some commands like `size` and `htop` revealed that 300 KB of memory out of the entire 11 MB is used for the search program itself and statically allocated variables, and 5.8 MB for shared libraries. Since we reuse some of DiskANN’s utilities for current AiSAQ implementation, better implementation of AiSAQ beam search may further reduce memory usage.

4.3 Query Search Time

Figure 3 shows recall-vs-latency plots of DiskANN and AiSAQ in each dataset. Since AiSAQ does not change the graph topology itself, recall@1 is identical to DiskANN in the same search condition. In some datasets like SIFT1M and KILT E5 22M, AiSAQ needs more blocks for a node chunk than DiskANN and the I/O request size for query search also becomes larger. Such datasets lead to I/O latency degradation compared with the same search conditions of DiskANN. Thanks to the I/O queueing system of SSDs and our optimization with shorter time for distance computation in CPU, however, the latency degradation is not critical. On the other hand, AiSAQ performs faster on SIFT1B because the I/O request sizes are the same $B_{AiSAQ} = B_{DiskANN} = 4$ KB.

We also conducted query search experiments for indices with various PQ vector sizes b_{PQ} to change the memory usage of DiskANN search. Figure 4 shows search latencies of DiskANN and AiSAQ search to achieve $> 95\%$ recall@1 of each b_{PQ} by memory usage. DiskANN searches configured to consume less RAM require slower latencies with large search list size L to ensure high recall@1. In contrast to DiskANN’s memory-latency tradeoff, AiSAQ keeps the same small memory usage regardless of PQ vector size b_{PQ} , which means

Table 3: Index load time (ms) of DiskANN and AiSAQ

	DiskANN	AiSAQ (ours)
SIFT1M	46.8	0.6
SIFT1B	16,437.4	0.6
KILT E5 22M	1,121.4	2.0

Table 4: KILT E5 2.2M subset index switch time (ms) of AiSAQ with and without PQ centroids reloading

	Index switch time (ms)
DiskANN	119.2
AiSAQ (PQ centroids reloading)	1.9
AiSAQ (shared PQ centroids)	0.3

AiSAQ can keep both of near-zero memory usage and millisecond-order search latencies in the high-recall area.

4.4 Index Switch

Table 3 shows the load time of DiskANN and AiSAQ indices before query search starts. While DiskANN’s load time increases with the dataset scale, AiSAQ keeps constantly short time since there is no need to load PQ vectors of the entire datasets. The load time of AiSAQ indices itself is the same order as its query search latencies. Higher-dimensional PQ centroids of KILT E5 22M brings longer time than the other datasets.

In applications of LLMs with RAG, it is possible that multiple datasets share the vector space (for example, images and documents are encoded to the same space by each encoder). In such situations, datasets are also able to share their PQ centroids. To obtain multiple datasets which can share the centroids, we devided KILT E5 22M dataset into 10 subsets, employing the PQ centroids generated by the whole 22M dataset. Then we built an index from each subset and searched across these indices with or without reloading the PQ centroids file. From table 4, sharing the PQ centroids, which only needs to load 4 KB metadata for an index, significantly reduced index switch time even compared to AiSAQ with reloading. This method is especially effective for high-dimensional datasets like KILT E5.

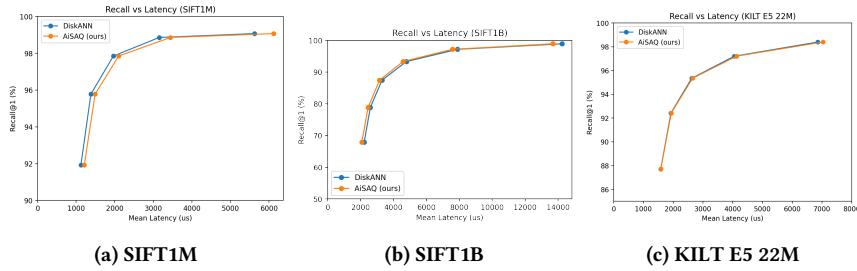


Figure 3: Recall vs latency graphs of DiskANN and AiSAQ

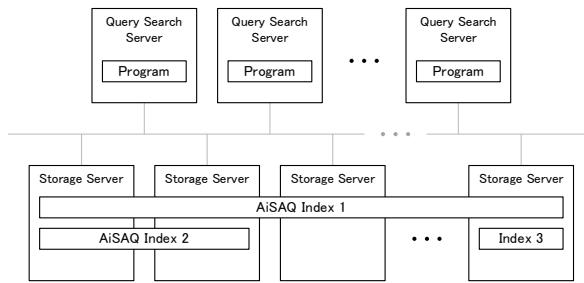


Figure 5: Concept of multiple-server AiSAQ system

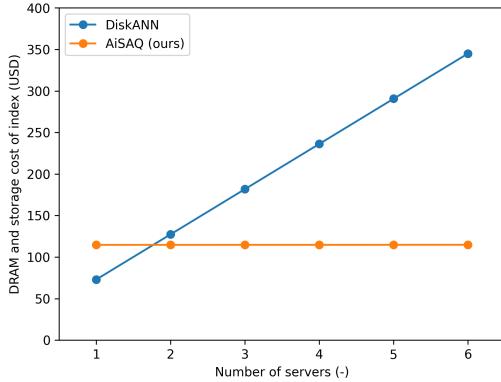


Figure 6: Cost estimation of DiskANN and AiSAQ (SIFT1B)

Table 5: DiskANN and AiSAQ comparison with 6 containers search (SIFT1B)

	DiskANN	AiSAQ
Total memory usage (MB)	183,405	66
Average index load time (ms)	14319.25	1.7
Estimated DRAM & SSD cost (USD)	344	103

4.5 Cost Analysis and Multiple-Server System

AiSAQ achieves similar high performance to DiskANN with extremely small memory usage and index load time, which results

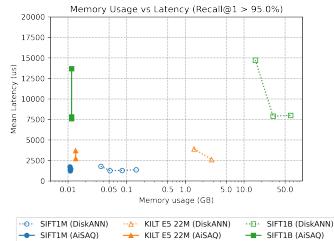


Figure 4: Latency vs memory usage of DiskANN and AiSAQ

in inflated SSD index size. To retain the PQ vector information, an AiSAQ index requires RNb_{PQ} bytes of storage, while DiskANN requires Nb_{PQ} bytes of DRAM. Maximum outdegree of the graph R is typically set around 60, while the cost per bit of DRAM is approximately 30 times of that of SSD (DRAM: 1.8 USD/GB vs SSD: 0.054 USD/GB, obtained from DRAMEXchange.com[1]). This means AiSAQ is not more cost effective than DiskANN in a single server.

However, it is possible that query search scales out with multiple servers like Figure 5 to handle increased search requests and larger indices. Such situations makes cost advantages to AiSAQ . Figure 6 shows estimated DRAM and SSD costs of DiskANN and AiSAQ . DiskANN search using n servers requires n times of DRAM capacity, while AiSAQ takes only the same capacity of storage as the case with a single server. AiSAQ overtakes DiskANN in terms of the resource cost in more than 2 search servers, which is reasonable for multiple-server search for larger-scale distributed ANNS system.

Table 5 shows experimental results and cost estimation of DiskANN and AiSAQ on multiple-server system. We used Lustre distributed filesystem with 3 storage servers for the index storage. AiSAQ search programs are executed on 6 Docker containers each regarded as a server on a single physical machine with an EPYC 7402P CPU and 224 GB of DRAM. All machines are connected by RDMA over 200 Gb ethernet. DiskANN with 6 containers consumes much more DRAM than AiSAQ , which can be the main factor of total cost. AiSAQ resulted in slower index load time than experiment with direct-attached SSD because of the slower latency over ethernet, but still reasonably fast time, while DiskANN did not achieve practical load time.

5 CONCLUSION AND FUTURE WORKS

In this paper, we proposed AiSAQ , a novel method of ANNS data offloading to storage. Our method achieved billion-scale ANNS with only \sim 10 MB memory usage without significant degradation of performance. AiSAQ will be applicable to all graph-based ANNS algorithms, and multiple-server system with lower costs.

In addition, reducing the index load time before query search enabled index switch between multiple large-scale dataset for various users' requests. That will enable LLMs with RAG to employ more simple index addition or filter search algorithms.

REFERENCES

- [1] Dramexchange - world leading dram and nand flash market research firm, with more than a decade of most authoritative database, 2024.

- [2] AMSALEG, L., AND JEGOU, H. Datasets for approximate nearest neighbor search, 2010.
- [3] ARYA, S., AND MOUNT, D. M. Approximate nearest neighbor queries in fixed dimensions. In *SODA* (1993), vol. 93, Citeseer, pp. 271–280.
- [4] CHASE, H. LangChain, Oct. 2022.
- [5] DILOCKER, E., VAN LUIJT, B., VOORBACH, B., HASAN, M. S., RODRIGUEZ, A., KULAWIAK, D. A., ANTAS, M., AND DUCKWORTH, P. Weaviate, 2022.
- [6] GOLLAPUDI, S., KARIA, N., SIVASHANKAR, V., KRISHNASWAMY, R., BEGWANI, N., RAZ, S., LIN, Y., ZHANG, Y., MAHAPATRO, N., SRINIVASAN, P., ET AL. Filtered-diskann: Graph algorithms for approximate nearest neighbor search with filters. In *Proceedings of the ACM Web Conference 2023* (2023), pp. 3406–3416.
- [7] JAYARAM SUBRAMANYA, S., DEVVRIT, F., SIMHADRI, H. V., KRISHNASWAMY, R., AND KADEKODI, R. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in Neural Information Processing Systems* 32 (2019).
- [8] JEGOU, H., DOUZE, M., AND SCHMID, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.
- [9] KARPUKHIN, V., OGUZ, B., MIN, S., WU, L., EDUNOV, S., CHEN, D., AND YIH, W. Dense passage retrieval for open-domain question answering. *CoRR abs/2004.04906* (2020).
- [10] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W.-T., ROCKTÄSCHEL, T., ET AL. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [11] MALKOV, Y. A., AND YASHUNIN, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.
- [12] PAN, Y., SUN, J., AND YU, H. Lm-diskann: Low memory footprint in disk-native dynamic graph-based ann indexing. In *2023 IEEE International Conference on Big Data (BigData)* (2023), IEEE, pp. 5987–5996.
- [13] PETRONI, F., PIKTUS, A., FAN, A., LEWIS, P., YAZDANI, M., DE CAO, N., THORNE, J., JERNITE, Y., KARPUKHIN, V., MAILLARD, J., PLACHOURAS, V., ROCKTÄSCHEL, T., AND RIEDEL, S. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, June 2021), Association for Computational Linguistics, pp. 2523–2544.
- [14] SIMHADRI, H., AUMÜLLER, M., BARANCHUK, D., DOUZE, M., INGBER, A., LIBERTY, E., LIU, F., AND WILLIAMS, G. Neurips’23 competition track: Big-ann, 2023.
- [15] SIMHADRI, H. V., KRISHNASWAMY, R., SRINIVASA, G., SUBRAMANYA, S. J., ANTONIJEVIC, A., PRYCE, D., KACZYNSKI, D., WILLIAMS, S., GOLLAPUDI, S., SIVASHANKAR, V., KARIA, N., SINGH, A., JAISWAL, S., MAHAPATRO, N., ADAMS, P., TOWER, B., AND PATEL, Y. DiskANN: Graph-structured Indices for Scalable, Fast, Fresh and Filtered Approximate Nearest Neighbor Search, 2023.
- [16] SINGH, A., SUBRAMANYA, S. J., KRISHNASWAMY, R., AND SIMHADRI, H. V. Freshdiskann: A fast and accurate graph-based ann index for streaming similarity search. *arXiv preprint arXiv:2105.09613* (2021).
- [17] WANG, J., YI, X., GUO, R., JIN, H., XU, P., LI, S., WANG, X., GUO, X., LI, C., XU, X., YU, K., YUAN, Y., ZOU, Y., LONG, J., CAI, Y., LI, Z., ZHANG, Z., MO, Y., GU, J., JIANG, R., WEI, Y., AND XIE, C. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data* (New York, NY, USA, 2021), SIGMOD ’21, Association for Computing Machinery, p. 2614–2627.
- [18] WANG, L., YANG, N., HUANG, X., JIAO, B., YANG, L., JIANG, D., MAJUMDER, R., AND WEI, F. Text embeddings by weakly-supervised contrastive pre-training, 2022.