General and Practical Tuning Method for Off-the-Shelf Graph-Based Index

SISAP Indexing Challenge Report by Team ***UTokyo***

Yutaro Oguri, Yusuke Matsui

1. Summary

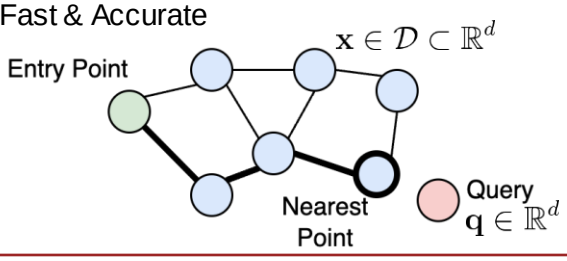
1. Indexing Challenge Task A

We got **2nd place** 🥈 in 10M and 30M track

- LAION5B [Schuhmann+, NeurIPS'22] subset
- Recall@10 ≧ 0.9

2. Graph-based index

- Moving forward from entry point to query
- Fast & Accurate



3. Performance Bottleneck

- Distance Computation (DC): $\arg\min_{\mathbf{x}\in\mathcal{D}} \|\mathbf{q}-\mathbf{x}\|_2$

4. Contribution

Proposed general and practical tuning

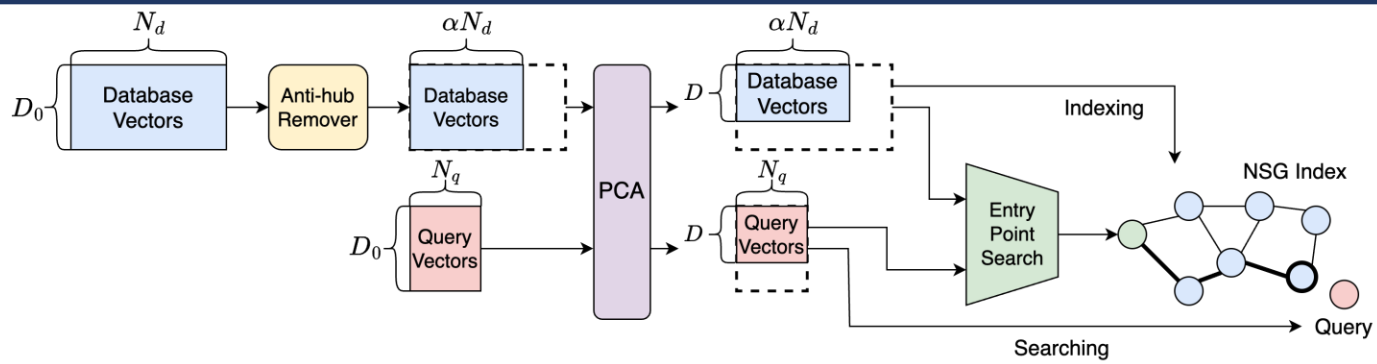
How to reduce the DCs

- Dimensionality Reduction using PCA
- Database (DB) Subsampling using Anti-hub Removal [Tanaka+, ICMR'21]
- Entry Point Selection using k-means

How to optimize them

- Black-box optimization with Optuna [Akiba+, KDD'19]

2. Method



- Dimensionality Reduction (PCA): Reduces dimensionality $D_0 \to D$.
- DB Subsampling (Removes insignificant points): Reduces size $N \to \alpha N$.
- Entry Point Selection: Partitions DB into k clusters. k centroids are candidates of entry point.
- Tuning D, α, k with black-box optimization algorithm
- Applicable to general off-the-shelf graph indexes

3. Evaluation

- Tuning NSG [Fu+, PVLDB'17] index
- Adopts the most efficient configuration with Recall@10 ≧ 0.9
- **Outperforms the vanilla NSG**

- Final ranking is evaluated with private query sets

	Recall@10(↑)	QPS [1/s] (↑)		
Size	Ours	Ours	Vanilla NSG [5]	Brute-force
300K	0.9208	$\mathbf{1.104 \times 10^5}$ (×**34.16**)	7.186×10^4 (×22.23)	3.232×10^3 (×1.0)
10M	0.9082	$\mathbf{3.822 \times 10^4}$ (×**1078**)	2.881×10^4 (×812.5)	35.46 (×1.0)
30M	0.9030	$\mathbf{3.010 \times 10^4}$ (×**1188**)	1.860×10^4 (×734.6)	25.32 (×1.0)

Team	Size	Query time (in *seconds*)	Team	Size	Query time (in *seconds*)
HSP	10M	0.34	HSP	30M	0.49
UTokyo	10M	0.49	UTokyo	30M	0.71
BASELINE-SearchGraph	10M	0.61	BASELINE-FAISSHNSW	30M	0.86
BASELINE-FAISSHNSW	10M	0.74	BASELINE-SearchGraph	30M	1.09