Probabilistic Routing for Graph-Based Approximate Nearest Neighbor Search

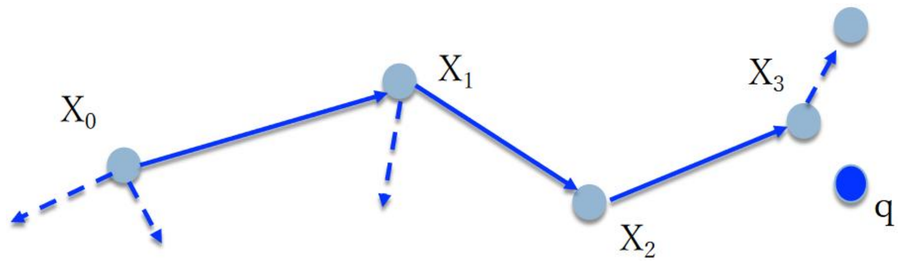
[1]Kejing Lu, [2]Chuan Xiao, [1]Yoshiharu Ishikawa
[1]Nagoya University, [2]Osaka University

1 Graph-based ANNS

◆Advantages of Graph-based ANNS

Compared with other ANNS techniques, graph-based ANNS achieves the best tradeoff between efficiency and accuracy.

◆A Simple Example of Graph-based ANNS



◆Limitations of Current Routing Strategies

Most of neighbors cannot be added into the node list. Scanning all the neighbors lacks effciency.

2 Partioned Extreme Order Statistics

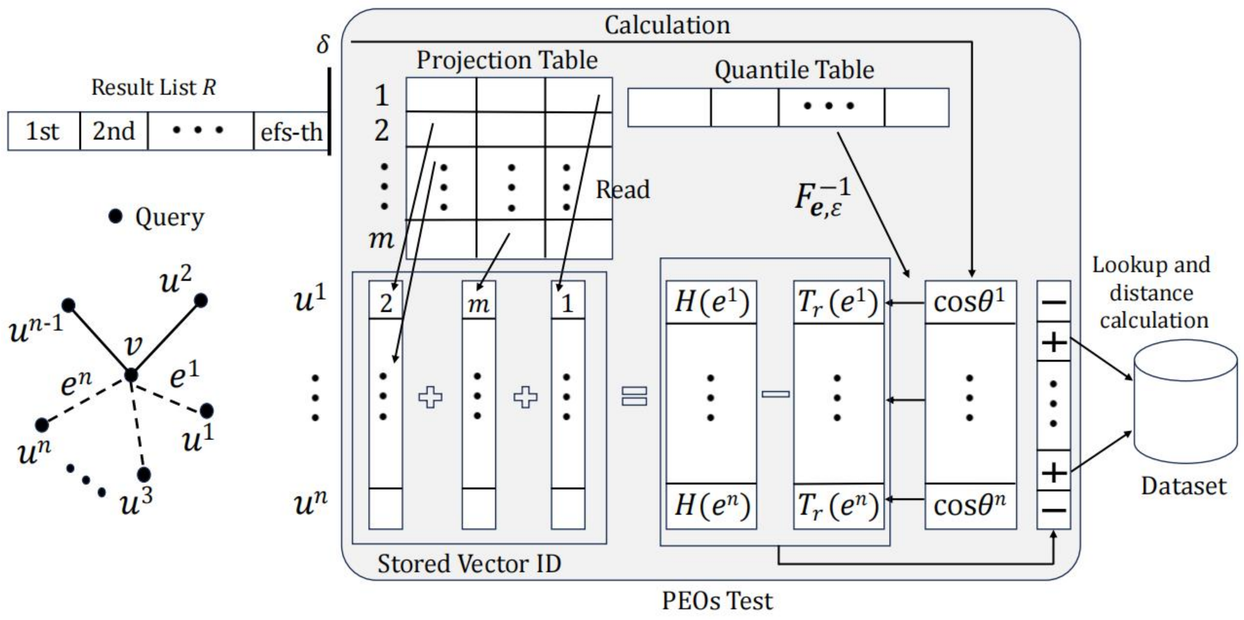
◆Core Idea

By using order statstics in each orthongal subspace and combining the extreme values, we can accurately estimate the angle of each data vector to the query vector.

◆Some Features

(1) We build a new hypothetical test to check each neighbor.
(2) The proposed test is friendly to SIMD acceleration.
(3) The poposed test can be combined with any graph structure.

◆An Illustration of PEOs Test



3 Theoretical Analysis

◆Key Parameters

m: the number of generated projected vectors.
L: the number of orthogonal subspaces.

◆Main Results

(Probabilistic Routing) Suppose that m is sufficiently large, PEOs test is (δ, 1-ε)-routing, where δ is the distance threshold and ε is the error rate.
(False Positives) As L increases moderately, the probability that false points pass the test decreases rapidly.

◆Some Observations

(1) L should not be too large or too small.
(2) Optimal L depends on the dimension of dataset.
(3) m=128 is enough for the pratical use.

4 VS Other Projection Techniques

Routing Test	#Projected Vectors	L	Code Length (Byte)
SimHash	$n = 64$	NA	8
RCEOs	$m > 428957$	$L = 1$	4
PEOs (opt)	$m = 128$	$L > 2.67$	4 (3+1)
PEOs	$m = 128$	$L > 2.69$	4 (3+1)

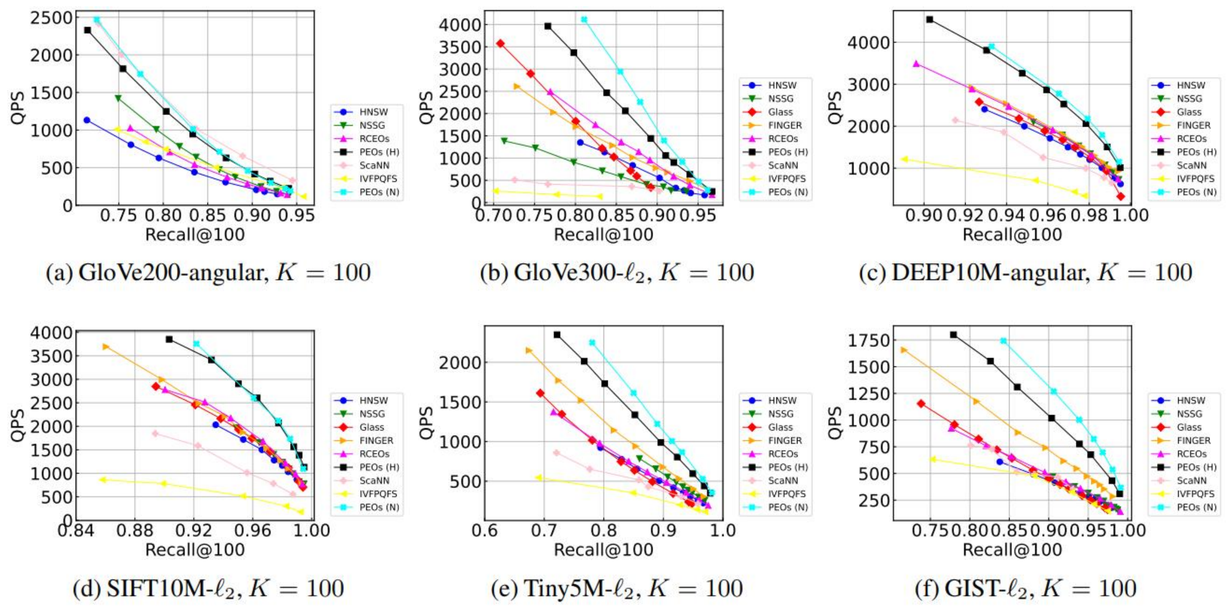
5 Experimental Results

◆Datasets and Benchmark Methods

Dataset	Size ($\lvert\mathcal{O}\rvert$)	Dim. (d)	Type	Metric
GloVe200	1,183,514	200	Text	angular
GloVe300	2,196,017	300	Text	ℓ_2
DEEP10M	9,990,000	96	Image	angular
SIFT10M	10,000,000	128	Image	ℓ_2
Tiny5M	5,000,000	384	Image	ℓ_2
GIST	1,000,000	960	Image	ℓ_2
DEEP100M	100,000,000	96	Image	angular

FINGER: simhash-based method
RCEOs: single-space-based method
ScaNN: Quantization -based method
...

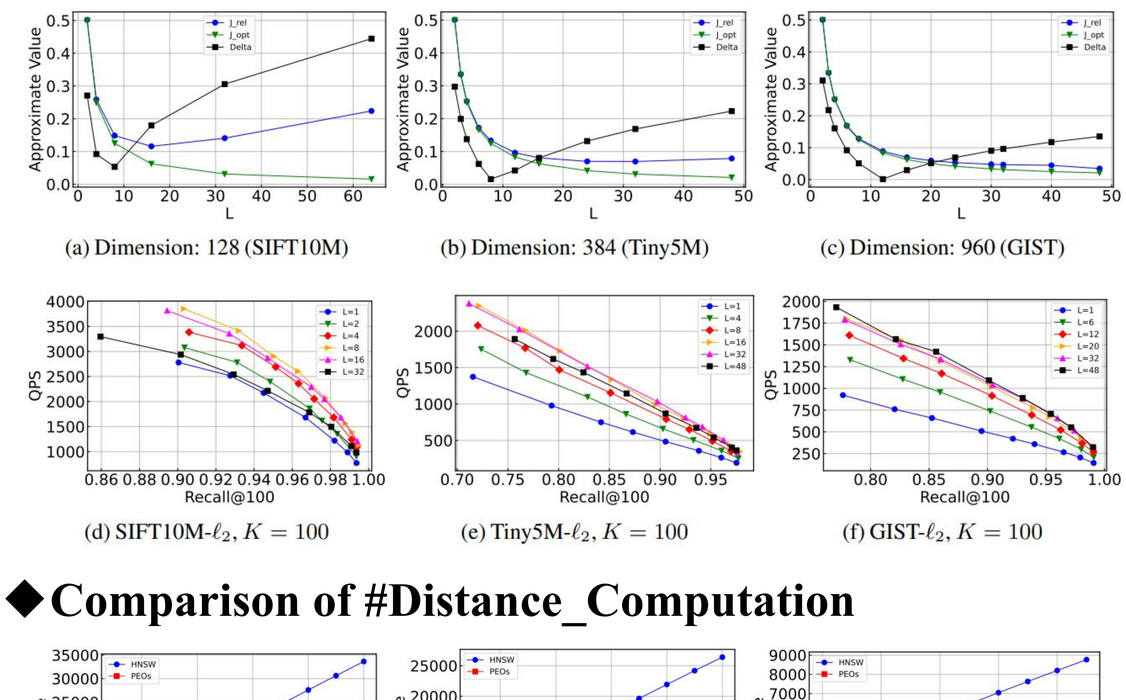
◆Performance Evaluation



(a) GloVe200-angular, $K = 100$ (b) GloVe300-ℓ_2, $K = 100$ (c) DEEP10M-angular, $K = 100$

(d) SIFT10M-ℓ_2, $K = 100$ (e) Tiny5M-ℓ_2, $K = 100$ (f) GIST-ℓ_2, $K = 100$

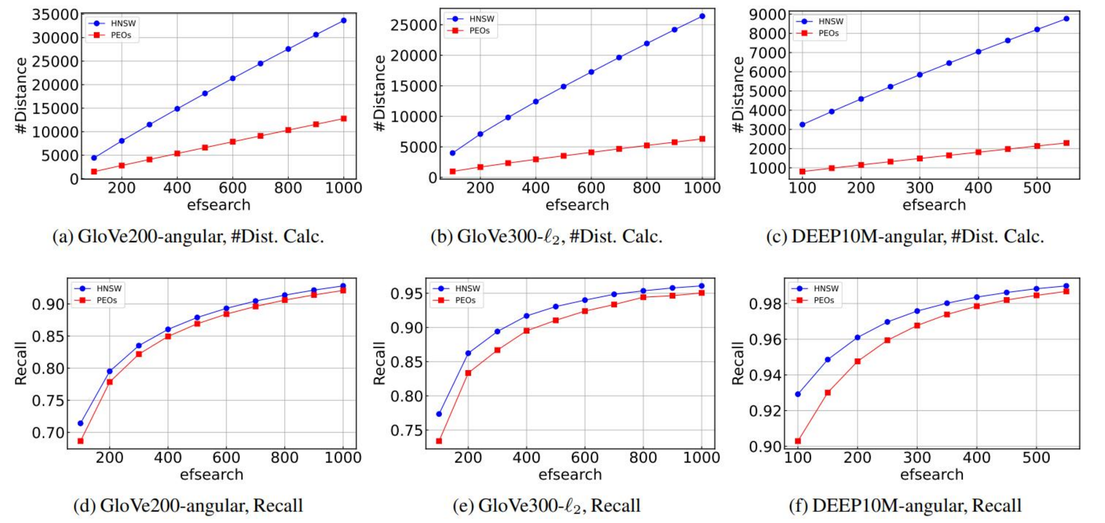
◆Effect of L



(a) Dimension: 128 (SIFT10M) (b) Dimension: 384 (Tiny5M) (c) Dimension: 960 (GIST)

(d) SIFT10M-ℓ_2, $K = 100$ (e) Tiny5M-ℓ_2, $K = 100$ (f) GIST-ℓ_2, $K = 100$

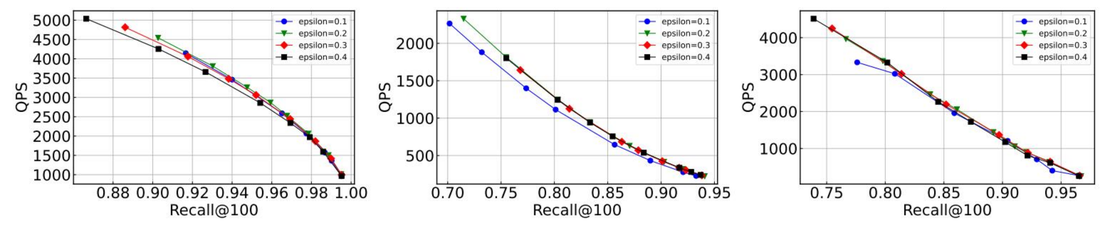
◆Comparison of #Distance_Computation



(a) GloVe200-angular, #Dist. Calc. (b) GloVe300-ℓ_2, #Dist. Calc. (c) DEEP10M-angular, #Dist. Calc.

(d) GloVe200-angular, Recall (e) GloVe300-ℓ_2, Recall (f) DEEP10M-angular, Recall

◆Effect of Epsilon



(a) DEEP10M-angular, $K = 100$ (b) GloVe200-angular, $K = 100$ (c) GloVe300-ℓ_2, $K = 100$