NEURAL INFORMATION PROCESSING SYSTEMS

Knowledge Distillation for High Dimensional Search Index

Zepu Lu[1,2], Jin Chen[3], Defu Lian[1,2*] , Zaixi Zhang[1,2], Yong Ge[4], Enhong Chen[1,2]
1.School of Computer Science and Technology, University of Science and Technology of China
2.State Key Laboratory of Cognitive Intelligence, Hefei, Anhui, China
3.University of Electronic Science and Technology of China 4.University of Arizona

中国科学技术大学 University of Science and Technology of China

电子科技大学 University of Electronic Science and Technology of China

THE UNIVERSITY OF ARIZONA

➢ Problem Background

- High Dimensional Index:
  - The advanced indexes, such as graph-based indexes, can yield more accurate results owing to their powerful expressiveness of the vector space
- Lightweight compressed index:
  - Notable for their substantial advantages in terms of low storage costs and efficient parallel processing. Existing learning methods for quantizers differ in whether there are explicit labels and whether they are related to queries.
    - Unsupervised algorithms only connected with document embeddings without the query information.
    - Relies on correlations with queries and documents but lacks explicit natural supervision signals.
    - Employs the query-dependent ground-truth labels to improve the retrieval performance.

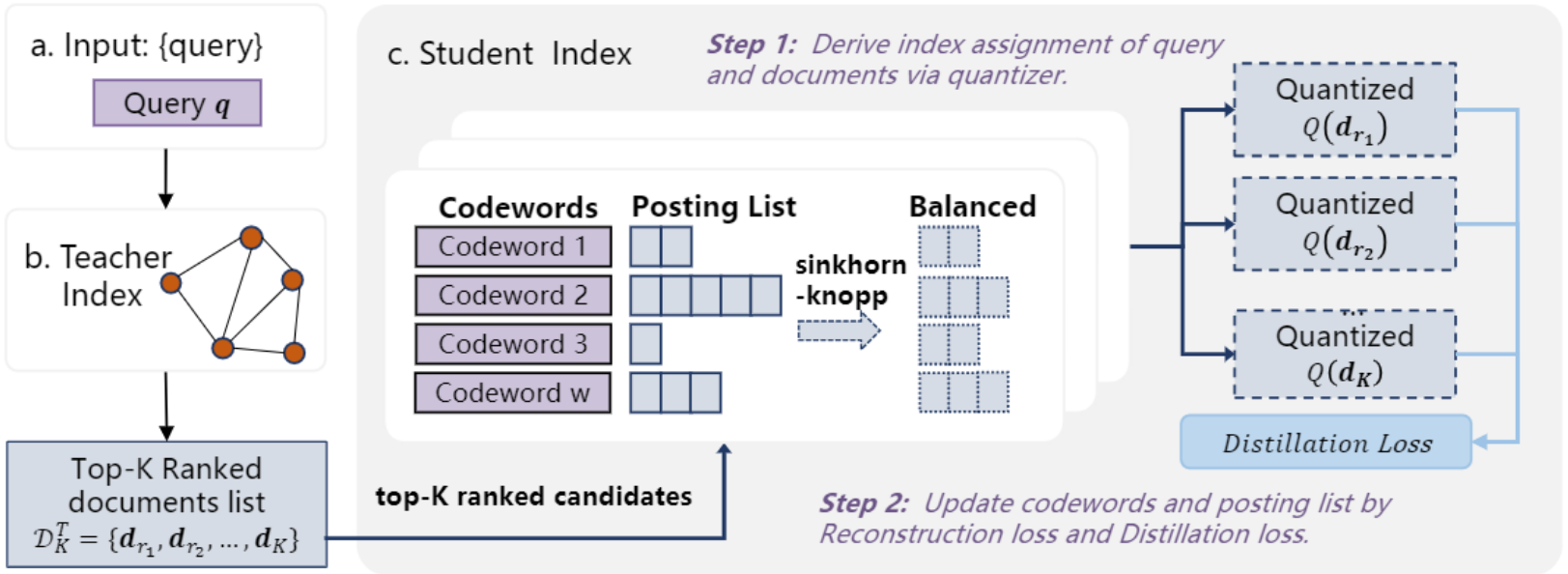
➢ Problem Background

- Issues of lightweight compressed index
  - The interaction label is only available in a small set of datasets and the expense of obtaining the ground-truth labels is particularly high, making the scenarios without any label information a more common case.

- Motivation
  - Learn the top-k nearest neighbors retrieved from the teacher models, which is label-free and takes the query information into consideration

➢ Proposed Solution: KDindex

- Knowledge Distillation for high dimensional search index (KDindex)



- Given the teacher index built on full-corpus, top-K ranked documents are returned for each query vector.
- Then, the query and corresponding documents are compressed by lightweight indexes and are ideally encoded with similar indexes, which are guided by the ***ranking-oriented distillation loss***.
- ***Two balance strategies***: Reconstruction loss and Sinkhorn-Knopp are used to avoid trivial solutions.

➢ Proposed Solution: KDindex

- Complexity Analysis

Methods	Initialization	Training	Indexing
KDindex (PQ)	$O(MWD)$	$O(MWD + MBW + MBWK)$	$O(NWD)$
KDindex (OPQ)	$O(MWD^2)$	$O(MWD^2 + MBW + MBWK)$	$O(NBW((D/B) + D^2))$
KDindex (AQ)	$O(MBWD)$	$O(MBWD + MBW + MBWK)$	$O(NBWD)$
HNSW	N/A	N/A	$O(ND \log N)$
ScaNN	N/A	$O(MWD + MBW + MBWK)$	$O(N(BK_p(D/B) + K_vD))$

The time complexity of KDindex, HNSW, and ScaNN. Denoted by D item embedding dimension, B the number of subspaces, W the number of centroids in each codebook, M the batch size (the number of queries in each batch), and K the number of neighbors. N is the number of items. As for ScaNN, K_v denotes the number of centroids in VQ and K_p in PQ.

No additional space complexity, acceptable time complexity.

➢ Experiments

- Overall performances
  - All KDindex methods show superior performances compared with according quantization methods, indicating the efficient and preferable utilization of the knowledge distillation.
  - KDindex methods can differentially training codebooks and index under distillation loss and two balance strategies, contributing to better model convergence.

Model	SIFT1M Recall@10	GIST1M Recall@10	MS MARCO Doc Recall@10	MS MARCO Passage Recall@10
PQ [24]	31.27±0.12	5.04±0.14	2.85±0.32	1.80±0.08
OPQ [18]	33.85±0.14	15.99±0.17	14.63±0.25	9.13±0.53
AQ [3]	35.48±0.34	19.32±0.19	16.73±0.57	10.18±0.69
DiffPQ [6]	29.01±0.24	4.43±0.29	4.17±0.26	2.53±0.06
DeepPQ [17]	25.02±0.50	4.39±0.18	6.47±0.33	8.43±0.72
GCD [25]	33.58±0.63	15.76±0.13	15.59±0.27	10.28±0.46
RepCONC [53]	33.59±0.32	15.32±0.46	16.07±0.31	10.37±0.37
PQVAE [49]	30.39±0.65	6.09±0.33	12.43±0.25	8.26±0.21
KDindex (PQ)	32.53±0.34	9.43±0.15	8.74±0.32	4.32±0.53
KDindex (OPQ)	34.77±0.47	17.32±0.17	16.70±0.28	10.66±0.69
KDindex (AQ)	37.30±0.17	21.33±0.23	18.93±0.76	11.19±0.26
	NDCG@10	NDCG@10	MRR@10	MRR@10
PQ	73.21±0.17	19.84±0.75	3.82±0.19	2.84±0.35
OPQ	75.76±0.09	49.10±0.74	34.75±0.14	29.06±0.49
AQ	77.82±0.42	60.33±0.63	38.52±1.02	33.52±0.58
DiffPQ	70.53±0.16	17.39±0.64	10.43±0.39	3.69±0.12
DeepPQ	65.80±0.75	16.60±0.71	14.90±0.18	28.33±0.73
GCD	75.49±0.23	48.82±0.58	38.89±0.29	34.01±0.62
RepCONC	75.47±0.52	47.69±0.79	39.03±0.73	34.27±0.59
PQVAE	69.84±0.27	23.48±0.42	34.47±0.92	27.91±0.74
KDindex (PQ)	73.90±0.59	30.62±0.58	17.64±0.13	6.42±0.59
KDindex (OPQ)	76.32±0.63	52.36±0.49	39.75±0.25	34.62±0.47
KDindex (AQ)	80.01±0.37	63.17±0.62	41.69±0.34	35.23±0.38

Knowledge Distillation for High Dimensional Search Index

Thank You