A Ablation Study

This document discusses the parameters of the CRANBERRY algorithm and confirms it's efficiency bottleneck. It can be mitigated using different parameters that lead to a speed-up of 51 % in comparison with the article[1]. We test just 100M dataset on Computer A, both described in the article.

Parameters used in the article try to minimise the number of vectors read from the secondary storage to be refined. Such parameters do not lead to the most efficient search due to contemporary NVMe SSD drives. The experiments below confirm that the filtering power of the Voronoi partitioning can be the bottleneck in the CRANBERRY since (almost) 1M vectors are then sorted according to the Hamming distance of sketches to the query sketch, and this sorting is slow.

Table 1: Varying search efficiency and effectiveness for different parameters of the CRANBERRY algorithm. Average values, 10NN queries on 100M dataset

Column 1	Column 2	Col. 3	Col. 4	Column 5	Col. 6	Col. 7
Limit on Voronoi cells size	Early termination: soft limit on accessed vectors	Sketch length	Average recall	Average number of accessed vectors	Σ time of 10,000 queries [s]	# queries per second
500,000	1,000	512	89.93%	1,015	1,353	7.39
550,000	1,000	512	90.09%	1,015	1,417	7.06
600,000	1,000	512	90.22%	1,015	1,442	6.94
700,000	1,000	512	90.52%	1,015	1,547	6.46
400,000	1,200	512	89.93%	1,217	1,571	6.37
800,000	800	512	89.87%	823	1,654	6.05
900,000	800	512	89.97%	823	1,842	5.43
800,000	1,000	512	90.70%	1,015	1,937	5.16
1,000,000	800	256	84.28%	848	1,942	5.15
1,000,000	800	384	88.40%	834	2,010	4.98
1,000,000	800	512	90.07%	823	2,139	4.67

Table 1 summarises the parameters and search times. All numbers are averages over the same 10,000 query vectors as used in the article. (1) The 1st column contains the maximum number of vectors in Voronoi cells that are further filtered by the CRANBERRY. (2) The soft limit on the refined vectors defines the early termination strategy and is published in the 2nd column. (3) The third column clarifies that the sketch length of 512 should not be decreased. (4) The goal of the article was to achieve an average search recall of at least 90 %. The measured recall is in the 4th column. (5) The actual average number of refined vectors is in the 5th column. (6) The 6th column provides the overall search-

[1] Mic, Vladimir; Sedmidubsky, Jan; Zezula, Pavel: *CRANBERRY: Memory-Effective Search in 100M High-Dimensional CLIP Vectors*, Similarity Search and Applications: 16th Int. Conf., SISAP 2023, Spain, Proceedings. Cham: Springer

ing time needed to evaluate all 10,000 queries. (7) The 7th column provides an average query throughput.

The last row in the table corresponds to the results presented in the article. The yellow row provides the lowest search time with a recall over 90 % and confirms that the more effective space partitioning and space pruning should be used to fully utilise the capabilities of sketches and the relational similarity.