Discussion Leader Papers

GUANGYU MENG

ACM Reference Format:
Guangyu Meng. 2022. Discussion Leader Papers . 1, 1 (March 2022), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Approximate nearest neighbor search (ANNS) constitutes an essential operation in many applications, including recommendation systems, information retrieval, and pattern recognition. In the past decade, graph-based ANNS algorithms have been the leading paradigm in this domain, with dozens of graph-based ANNS algorithms proposed. However, the graph-based ANNS algorithms suffer from long graph construction, graph search time, and large memory consumption.
I select three papers that optimize graph-based ANNS in algorithm and hardware manners.

2 EFFICIENT AND ROBUST APPROXIMATE NEAREST NEIGHBOR SEARCH USING HIERARCHICAL NAVIGABLE SMALL WORLD GRAPHS [2]

This hierarchical navigable small world graphs(HNSW) work is accepted by the 2018 TPAMI journal, and it is the pioneer for the graph-based ANNS. HNSW overcomes the search complexity from poly-logarithmic complexity into logarithmic complexity in the graph-based ANNS. It generates a hierarchical graph and fixes the upper bound of each vertex's number of neighbors, thereby allowing a logarithmic complexity scaling of search. Its basic idea is to separate neighbors to different levels according to the distance scale, and the search is an iterative process from top to bottom. For an inserted point, HNSW not only selects its nearest neighbors, but also considers the distribution of neighbors. HNSW has been deployed in various applications because of its unprecedented superiority.

3 HM-ANN: EFFICIENT BILLION-POINT NEAREST NEIGHBOR SEARCH ON HETEROGENEOUS MEMORY [3]

Efficient billion-scale approximate nearest neighbor search (ANNS) has become a significant research problem, inspired by the needs of machine learning based applications. Among these ANNS methods , it has been demonstrated that similarity graphs, such as Hierarchical Navigable Small World (HNSW) [2], obtain superior performance relative to tree structure based, locality sensitive hashing (LSH) based, and inverted multi-index (IMI) based approaches on most public benchmark dataset. However, one major limitation of existing similarity graphs is that they are very memory consuming and easily run out of memory with a few hundred millions of vectors.
The emergence of heterogeneous memory (HM) brings opportunities to largely increase memory capacity and break the above trade-off: Using HM, billions of data points can be placed in main memory on a single machine

Author's address: Guangyu Meng.

XXXX-XXXX/2022/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

without using any data compression. In this paper, the authors present a novel graph-based similarity search algorithm called HM-ANN, which takes both memory and data heterogeneity into consideration and enables billion-scale similarity search on a single node.
This paper is accepted by the NeurIPS 20 conference, and some solutions for the NeurIPS 21 competition track "Billion-Scale Approximate Nearest Neighbor Search Challenge" are based on this paradigm.

4 SPANN: HIGHLY-EFFICIENT BILLION-SCALE APPROXIMATE NEAREST NEIGHBORHOOD SEARCH [1]

This paper is from Microsoft(the Bing team) and was accepted as the spotlight paper for the NeurIPS 2021 conference. In this paper, the authors study the approximate nearest neighbor search (ANNS) problem and develop an inverted index based algorithm using both memory and disk in the searching to reduce the required amount of memory. The proposed algorithm first partitions the vectors into clusters and then stores in the memory the centroid of each cluster and the disk address of the vectors belonging to the cluster, or called posting lists. To limit the length of posting lists (the number of vectors in a cluster), the algorithm performs the clustering with an additional constraint of balancing the length of each posting list. But the algorithm performs the clustering in a hierarchical way so that the computational cost can be improved. It further replaces a centroid by the closest vector and another index [10] to speed up the search. It also assigns boundary vectors, those close to multiple clusters, to the multi-clusters to reduce the number of posting lists accessed. Finally, during the query, only posting lists that are sufficiently close to the query compared to the closest posting list will be searched. Experimental evaluations have been conducted on two datasets of one billion to demonstrate the superior performance.

REFERENCES

[1] Qi Chen et al. “SPANN: Highly-efficient Billion-scale Approximate Nearest Neighborhood Search”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[2] Yu A Malkov and Dmitry A Yashunin. “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.4 (2018), pp. 824–836.

[3] Jie Ren, Minjia Zhang, and Dong Li. “Hm-ann: Efficient billion-point nearest neighbor search on heterogeneous memory”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10672–10684.