Parameter Tuning in Graph-Based Approximate Nearest Neighbor Methods

The recent growth of graph-based approaches in Approximate Nearest Neighbor (ANN) search has raised numerous questions about parameter tuning in these methods. Although different algorithms may use various names for their parameters, they generally serve similar functions across different methods.

Key Parameters

1. Maximum Outdegree

- **Definition**: Limits the size of the neighborhood list for each node, controlling the number of connections per node.
- **Influence on Indexing Footprint and Size**: Affects the space occupied by the neighbor IDs of each node, influencing the overall index size.
- **Impact on Search Performance**: Balances the trade-off between index size and search efficiency.

2. Beam Width During Construction

- **Definition**: Used in methods that build the neighbor list based on the results of a beam search on the graph (or a partial graph in incremental builds).
- **Influence on Indexing Footprint and Size**: Has minimal impact on index size.
- **Impact on Search Quality and Performance**: Significantly affects search quality and performance by controlling the quality of each node's neighbors, and consequently, the quality of the graph itself.

Secondary Parameters

Methods like NSG, VAMANA, ELPIS, and SSG introduce additional parameters:

- **Range** : Use in both NSG, and VAMANA, where the used beam width is small, but the candidate neighbors considered during pruning are the list of visited nodes; this parameters defines the candidate set size.
- **Alpha** : Pruning parameter for RRND and MOND
- **Leaf Size** : used in ELPIS, this parameters defines the maximum size of clusters.
- **Number of Seed Selection Trees** : appears in methods that uses tree structures such as KD tree, BKtree or VP tree for seed selections or init the node neighbors.
- **Number of Clusters** used in methods where the tree are not considered and directly clustering of the data is used.

These parameters are generally considered after tuning the maximum outdegree and beam width, as they have a secondary impact on performance.

HNSW: A Benchmark Method

The Hierarchical Navigable Small World (HNSW) algorithm is currently considered the state-of-the-art method for graph-based ANN due to:

- **Efficiency**: Offers high-performance search capabilities.
- **Ease of Parameter Tuning**: Limits its parameters to two primary ones:
  - **Maximum Outdegree (M)**: Controls the number of connections per node in hierarchical layers (M for higher layers and $M \times 2$ for the base layer).
  - **Beam Width (*efConstruction*)**: Adjusts the number of candidate neighbors during graph construction.

Advantages of HNSW Parameters

- **Simplicity**: Only two parameters to tune, making it user-friendly.
- **Performance**: Proper tuning can lead to high recall rates (e.g., 99% recall) with efficient search performance.

Performance Impact Illustration

The figure below demonstrates how changing these parameters influences search performance, focusing on scenarios requiring high recall (e.g., 99%).

n Figure 1 above (where green means best and red means worst), we compare search performance at 0.99 recall for 100 queries on HNSW indices built

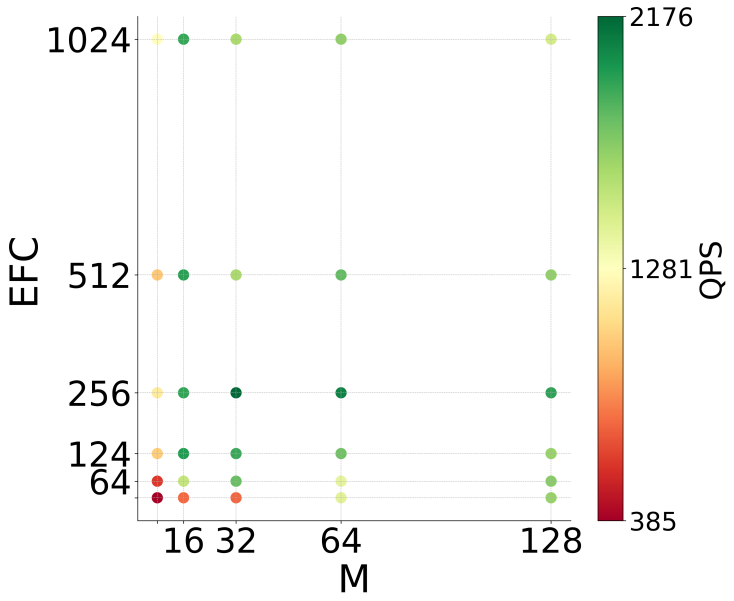

Figure 1: Indexing Parameters Impact on Search Efficiency at 0.99 recall

with different outdegrees (M, ranging from 8 to 128) and beam widths (EFC, ranging from 32 to 1024). At low M (8), high recall requires a high EFC during search. Increasing EFC during indexing improves the quality of neighborhood candidates, which is critical for diversification approaches like HNSW, ELPIS, or VAMANA. However, beyond EFC=128, search efficiency does not improve further, as more neighbors are required per node. Increasing both M and EFC to their maximums is not advisable, as high outdegree leads to large neighborhood sets, rendering Neighbor Diversification pruning ineffective. Besides, high values of M or EFC increase indexing time; for instance, M=128 and EFC=1024 is 28x more expensive than M=8, EFC=32, and 23x more expensive than M=32, EFC=256. Increasing both parameters increases indexing time due to larger beam widths during insertion and more neighbors to compare during routing. Tuning both parameters can be time-consuming, especially on large datasets. Therefore, for HNSW and other methods that use a maximum outdegree and beam width as indexing parameters, we tune the parameters by selecting the best values from a defined range. This range is chosen based on literature and/or from our experiments on previous similar datasets with comparable hardness and/or size.