Task: The objective is to create an approximate K-NN graph using a set of high-dimensional vectors where each vertex is linked to its approximate k nearest neighbors based on the Euclidean distance.

Dataset:

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of vectors</th>
<th>dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turing</td>
<td>10,000,000</td>
<td>100</td>
</tr>
</tbody>
</table>

Vectors are Bing queries encoded by Turing AGI v5.

Input: A set of vectors

Output: Approximate KNN Graph (K=100)

Evaluation Metric: \( \text{Recall} = \frac{\text{ground truth (100--NN neighbors)}}{100} \)

Hardware Conditions: Azure Standard F32s_v2, 32C64G

Time Limit: 30 minutes + 60 seconds (reprozip overhead)

Solution Overview

Unfortunately, the curse of dimensionality frequently causes hubs to appear in high-dimensional space, which we also observed in the Turing Dataset during our analysis. This causes the reverse-NN set to be truncated, thereby saving memory budget, but also undermining neighbor discovery through reverse-NN paths.

With Slack-Threshold:

This method is based on two key principles:

- A neighbor of a neighbor is also likely to be a neighbor. (from NN-Descent)
- A closer neighbor could be found by taking a slight detour. (Inspired by the RNG property [2])

With Slack-Threshold, nodes with relatively far distance from \( P \) are more likely to include \( P \) into its neighbor checking phase. (Node X)

It filters out Node Y. However, by applying local join and relying on principle 1 of NN-Descent, we ensure that local close nodes are detected and well-connected. This approach increases the likelihood of reaching nodes that would otherwise require detours, while also preserving strong local connections.

Low-Level Design:
- The previously mentioned structure will be dynamically maintained to reduce memory footprint.
- SIMD with AVX-512 is used with aligned data for accelerating L2 distance computation.
- Prefetching has been effectively integrated into the computational checking progress.

Results

Results are selected from submissions in the contest.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recall</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turing-10M</td>
<td>0.981</td>
<td>1847</td>
</tr>
<tr>
<td>Turing-10M</td>
<td>0.976</td>
<td>1654</td>
</tr>
<tr>
<td>Turing-10M</td>
<td>0.954</td>
<td>1300</td>
</tr>
</tbody>
</table>

References
