SISAP Indexing Challenge 2025 Solution for Task 2 Using Root Join

Benjamin Bustos[0000-0002-3955-361X] and Jiale Chen[0009-0006-0942-3693] , Universidad de Chile
IMFD, Department of Computer Science, University of Chile, Chile
bebustos@dcc.uchile.cl, jiale.chen@ug.uchile.cl

SISAP Indexing Challenge 2025 - Task 2

Task 2 from SISAP Indexing Challenge 2025 [4,5] consists in the construction of a k-NN graph (self-similarity join) under hardware restrictions. For this task, the graph construction requires using k=15 with 384-D vectors and the dataset size is around 3 million.

Challenge constraints: Execution in a Linux container with 8 virtual CPUs, 16GB RAM and a limit computation time of 12 hours.

Our Solution

We propose a solution based on Root Join [1], an approximated algorithm for computing a self-similarity join that uses $\Theta(n^{3/2})$ distance computations , with n the size of the dataset. We added some pre-processing steps to improve its performance under the conditions of the Challenge. The main steps of Root Join are:

1- **Partition Strategy**

- Select $\sqrt{n}$ random points as centers.
- Each center forms a group of maximum size $c\sqrt{n}$ (c constant).
- Each element of the dataset is assigned to the group with the closest center, that has available space.

2- **Computing the Approximated k-NN Self-similarity Join**

- For each element s in a group, the algorithm computes a “target set” with the elements from the same group and the elements of next closest group. If necessary, the target set is expanded until reaching a size of at least k.
- The algorithm finds the k nearest neghbors of s within the target set.
- The algorithm returns a set of pairs *(element, list of nearest neighbors).*

Main Modifications of Root Join

1- **Group Uniformity**: In the original Root Join, if $c > 1$, some groups may be completely full, while others may be almost empty.

To avoid this problem, we consider $c = 1$. This guarantees that there are at least $\lfloor\sqrt{n}\rfloor$ elements in the $\lfloor\sqrt{n}\rfloor$ existing groups. The remaining $\mathrm{x} = \mathrm{n} - \lfloor\sqrt{n}\rfloor\lfloor\sqrt{n}\rfloor$ elements are uniformly distributed among the groups. With this process, we can solve the self-similarity join for $\mathrm{k} < \lfloor\sqrt{n}\rfloor$ even if considering as target set just the original group of an element.

2- **Dimensionality Reduction:** We use Principal Component Analysis [2] (PCA), which is scalable for large dataset [3], for efficient distance computation.

3- **Increasing the Target Set:** In our implementation, we consider as target set the elements of the group of s, and we expand it with the two groups with the closest centers to s.

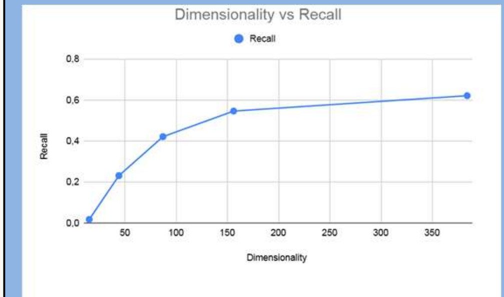

Fig. 1. Dimensionality vs Recall

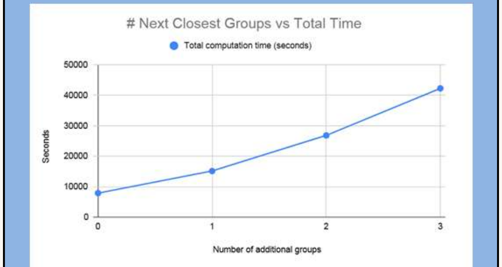

Fig. 3. Additional groups vs Total time

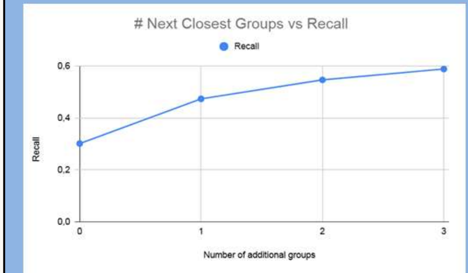

Fig. 2. Additional groups vs Recall

Acknowledgments.

This work was funded by ANID - Millennium Science Initiative Program - Code ICN17_002. We thank Sebastian Ferrada and Diego Arroyuelo for the discussions about the improvements to Root Join.

References

1- Ferrada, S., Bustos, B., Reyes, N.: An efficient algorithm for approximated self-similarity joins in metric spaces. Information Systems 91, 101510(2020). https://doi.org/10.1016/j.is.2020.101510 , https://www.sciencedirect.com/science/article/pii/S0306437920300211.

2- Gewers, F.L., Ferreira, G.R., Arruda, H.F.D., Silva, F.N., Comin, C.H., Amancio,D.R., Costa, L.D.F.: Principal component analysis: A natural approach to data exploration. ACM Comput. Surv. 54(4) (May 2021). https://doi.org/10.1145/3447755.

3- McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform manifold approximation and projection for dimension reduction, https://umap-learn.readthedocs.io/en/latest/benchmarking.html, (Last accessed 2025/06/01).

4- Tellez, E.S., Chavez, E., Aumüller, M., Mic, V.: Overview of the SISAP 2025 Indexing Challenge. In: Similarity Search and Applications: 18th International Conference, SISAP 2025, October 1st-3rd, Proceedings. vol. 16134. Springer-Verlag, Berlin,Heidelberg (2025).

5- Téllez, E.S., Chavez, E.L., Aumüller, M., Mic, V.: SISAP Indexing Challenge2025, https://sisap-challenges.github.io/2025/index.html, (Last accessed 2025/06/01).