Optimizing Vector Database Retrieval through Adaptive Hierarchical Clustering

Kathan Piyushkumar Bhavsar   
*University of Windsor*  
bhavsa85@uwindsor.ca

Kalp Jigneshbhai Kalani  
*University of Windsor*kalanik@uwindsor.ca Dev Pinakinbhai Patel  
*University of Windsor*patel41b@uwindsor.ca

Kush Janak Patel  
*University of Windsor*patel9ib@uwindsor.ca Mohammad Arsh Vahora  
*University of Windsor*vahora4@uwindsor.ca

Hetanshu Hemant Patel  
*University of Windsor*patel8qb@uwindsor.ca

*Abstract*— We propose an improved approach for vector similarity search optimization based on an enhanced hierarchical clustering system. The current solutions have three main flaws, such as efficiency in database restructuring, feature preservation, and multi-cluster membership handling. Our proposed system introduces adaptive parameters for controlled restructuring, weighted centroid calculation, and a skip algorithm for multi-cluster elements in order to reduce the search complexity with quality maintenance of results and address maintenance challenges in the vector databases. It integrates three key novel solutions: adaptive restructuring mechanism, context-preserving centroid computation, and skip algorithm for multi-cluster management; all together, these provide a more efficient and maintainable vector database system.

Keywords — vector database, hierarchical clustering, similarity search

# Problem Description

Modern applications ranging from recommendation systems to semantic search engines have been integrating vector databases. These databases store and process high-dimensional vectors representing complex data, such as images, text embeddings, and user behaviour patterns. However, current implementations are afflicted with several critical challenges that limit their effectiveness, especially considering the ever-increasing scale and complexity of applications [2].

The challenges are mainly centered on search efficiency, where the need to compare the query vectors with every first-level centroid creates a huge computational bottleneck. Since traditional methods require exhaustive comparisons even for preliminary clustering steps, these have linear time complexity concerning database size [4]. This problem scales up with the scaling up of databases, hence resulting in poor performance in large-scale applications. This leads to the fact that in real-world datasets, the query time could be above the acceptable threshold of interactive applications, particularly those systems relying on real-time answers [3].

Another major challenge is contributed by the maintenance overhead, as most of the existing hierarchical methods have to be remade in their entirety in case of new insertions. The addition of new vectors in current systems is many times achieved at the cost of going through several layers of hierarchy again for optimal clustering, which has significant computational cost and takes down the system from availability [5]. The problem is worse in dynamic systems, where updates should be very frequent, since every update operation can start expensive reorganization processes in the whole database structure [5].

At higher levels of granularity, feature dilution occurs due to simple averaging while calculating centroids representative of meaningful features and context of the original vectors. Moving deeper in the hierarchy, each higher level of abstraction loses more of the distinctive characteristics that make the original vectors meaningful; hence, potential loss in precision due to cluster overlap may deteriorate the quality of the search results [1]. This effect is amplified for applications in which small-scale vector differences carry important semantics, such as semantic search or recommendation systems, where subtle relationships between items must be preserved [3].

Current systems also have difficulty dealing with multi-cluster membership, and often one has to make arbitrary assignments for vectors that rightfully belong to multiple clusters. Many real-world entities naturally share characteristics with multiple groups, but traditional clustering approaches typically force each vector into a single cluster. This may seriously affect retrieval accuracy and is not representative of the natural relationships in data [4]. In such cases, for instance, the content recommendation system would classify a movie either under "action" or "comedy", while such a movie might belong to both categories. In this case, the recommendations are either incomplete or totally wrong [3].

# Motivation

In this regard, exponential growth in size and utilization of vector databases demands further scalable search methods. Applications keep generating more and processing increased volumes of vector data, thereby giving rise to emergent needs concerning the efficient mechanism for searching. Most of the current solutions cannot maintain acceptable performance with growing data [2]. There lies great potential for optimization in the computational overhead during database updates while preserving cluster quality. This creates the bottleneck for executing restructuring operations at high costs for real-time applications in current systems, particularly dynamic ones that do need updates quite frequently [5].

Preserving vector semantics across hierarchical levels is central to achieving meaningful and precise search results. The loss of semantic information in current systems leads to decreased search precision and reduced overall system effectiveness [1]. The accommodation of natural multi-cluster memberships is of growing importance for many real-world applications. Many contemporary applications require vectors to belong to multiple clusters, making this capability essential in truly representing the data and performing accurate retrieval [4].

# Solution

Our proposed solution meets these challenges by means of three key innovations.

Adaptive Restructuring System introduces user-controlled parameters in order to give the power of fine-tuning to database maintenance. It basically allows the setting of a tree height to be configured as per customized hierarchical depth; hence, a user can make a trade-off between the speed of search versus memory use. The system continuously monitors cluster sizes through a rechecking factor that triggers reorganization only when necessary. It also applies selective tree reorganization using new cluster factors so that the restructuring operations will affect only the affected clusters and not the whole database. This has massively reduced the maintenance overhead while still enabling users to change parameters w.r.t. their performance requirements and needs for stability [5].

Context-Preserving Centroid Calculation introduces an efficient way of maintaining semantic relevance along the hierarchy. This method applies the inverse multiplying weights in the calculation and allows dynamic weight adaptation concerning the similarity distances between vectors. It gives more importance to the contribution at the border for giving a better edge to each cluster and retaining semantic boundaries [8]. Iteratively, the position of centroids is optimized in order for them to represent their cluster best. This sophisticated approach maintains semantic relevance at higher levels of hierarchy, hence giving more accurate search results and better cluster representation across the entire database structure [4].

The Skip Algorithm for Multi-Cluster Management proposes a two-phase clustering scheme to manage complex vector relationships. The system identifies and temporarily skips the vectors showing significant similarity in multiple clusters without forcing them to immediate assignment during the first phase of clustering. The secondary clustering process of these skipped vectors analyzes the detailed relation with the previous clusters. Finally, the system will perform multi-cluster assignment on vectors that may have remained ambiguous, allowing the vectors to point to multiple clusters to which they relate [8]. This organized way of treating vectors that naturally belong to several clusters significantly enhances the overall accuracy of the system.

Together, these components present a complete package for tackling core problems that come with most state-of-the-art vector database solutions. Equipped with an adaptive restructuring system, context-preserving calculation method, and with multi-cluster sophisticated handling, the proposed system performs with increased effectiveness and quality while being able to provide efficiency during real-life applicability.

This is a proposed solution that takes advantage of combining advanced toolchains and technologies in order to achieve better performance for functionalities within vector databases. The Python-based frameworks include an adaptive restructuring system, a context-preserving centroid computation, and a skip algorithm for managing multiple clusters-all built on NumPy for vector operations and SciPy for clustering algorithms. Machine learning libraries such as PyTorch are further used to dynamically adjust weights and optimize centroids so that semantic relevance can be maintained in high-dimensional data.

##### References

1. P. S. Nishant, S. Mehrotra, P. R. Sree and P. Srikanth, "Hierarchical clustering based Intelligent Information Retrieval approach," *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2020, pp. 862-866. <https://ieeexplore.ieee.org/abstract/document/9214231?casa_token=JTIN4_SZ4pYAAAAA:Vb2i77RGDiMJm4Vusl1_5HWvx2nCXXtiuQzFkA5feLqycuqv_-gu069sZd5C3gkVKaQXIVYDdf0>
2. C. Chen, C. Jin, Y. Zhang, S. Podolsky, C. Wu, S. Wang, E. Hanson, Z. Sun, R. Walzer, and J. Wang. 2024. “SingleStore-V: An Integrated Vector Database System in SingleStore.” Proc. VLDB Endow. 17, 12 August 2024, 3772–3785. <https://doi.org/10.14778/3685800.3685805>
3. S. Kale, “From Text to Recommendations: How Vector Databases are Revolutionizing Personalized Content Delivery,” 2024. International Journal for Research in Applied Science & Engineering Technology (IJRASET). vol. 12, April 2024, ISSN: 2321-9653. <https://www.researchgate.net/profile/Satyanand-Kale/publication/380208131_From_Text_to_Recommendations_How_Vector_Databases_are_Revolutionizing_Personalized_Content_Delivery/links/6697340d8dca9f441b82f3f2/From-Text-to-Recommendations-How-Vector-Databases-are-Revolutionizing-Personalized-Content-Delivery.pdf>
4. S. Naumov, G. Yaroslavtsev, and D. Avdiukhin, “Objective-Based Hierarchical Clustering of Deep Embedding Vectors”, *AAAI*, vol. 35, no. 10, pp. 9055-9063, May 2021. <https://ojs.aaai.org/index.php/AAAI/article/view/17094/16901>
5. A. Ponomarenko, Y. Malkov, A. Logvinov, and V. Krylov. "Approximate nearest neighbor search small world approach." In International Conference on Information and Communication Technologies & Applications, vol. 17. 2011. <https://www.iiis.org/CDs2011/CD2011IDI/ICTA_2011/PapersPdf/CT175ON.pdf>
6. Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824-836, 1 April 2020 <https://ieeexplore.ieee.org/abstract/document/8594636>
7. N. Monath, K. A. Dubey, G. Guruganesh, M. Zaheer, A. Ahmed, A. McCallum, G. Mergen, M. Najork, M. Terzihan, B. Tjanaka, Y. Wang, Y. Wu. 2021. Scalable Hierarchical Agglomerative Clustering. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21), August 14–18, 2021 <https://dl.acm.org/doi/pdf/10.1145/3447548.3467404>
8. James C. Bezdek, Robert Ehrlich, William Full. “FCM: The fuzzy c-means clustering algorithm”. Computers & Geosciences, Volume 10, Issues 2–3, 1984, Pages 191-203, ISSN 0098-3004, <https://doi.org/10.1016/0098-3004(84)90020-7>.