**Project Proposal: Optimizing Vector Database Retrieval through Adaptive Hierarchical Clustering**

**Abstract**

This project proposes an innovative approach to optimize vector similarity search in vector databases through an enhanced hierarchical clustering system. While existing solutions employ hierarchical structures, they face challenges with database restructuring, feature dilution, and multi-cluster membership. Our proposed system introduces adaptive parameters for controlled restructuring, weighted centroid calculations to preserve feature context, and a novel skip algorithm for handling multi-cluster elements. This solution aims to significantly reduce search space while maintaining search quality and addressing the structural maintenance challenges of current vector databases.

**Problem Description**

Vector databases have become crucial in modern applications, from recommendation systems to semantic search. However, current implementations face several critical challenges:

1. Search Efficiency: The necessity to compare query vectors against all first-level centroids creates a computational bottleneck, especially as databases scale.
2. Maintenance Overhead: Existing hierarchical implementations require complete tree restructuring upon new insertions, making them computationally expensive and potentially disrupting system availability.
3. Feature Dilution: As centroids are calculated at higher levels through simple averaging, the meaningful features and context of the original vectors become increasingly diluted, leading to potential precision loss and cluster overlap.
4. Multi-Cluster Membership: Current systems struggle to efficiently handle vectors that legitimately belong to multiple clusters, often forcing arbitrary assignments that can impact retrieval accuracy.

**Motivation**

The exponential growth in vector database applications demands more efficient and maintainable solutions. Our research is motivated by:

1. The need for more scalable vector search methods that can handle large-scale databases without compromising on search speed or accuracy.
2. The opportunity to reduce computational overhead during database updates while maintaining cluster quality.
3. The potential to preserve vector semantics across hierarchical levels, ensuring more meaningful and accurate search results.
4. The importance of accommodating natural multi-cluster memberships in real-world applications.

**Solution**

Our proposed solution introduces three key innovations:

**1. Adaptive Restructuring System**

* Implementation of user-controlled parameters:
  + Tree height configuration for customized hierarchical depth
  + Rechecking cluster size factor for localized cluster stability monitoring
  + Restructure new cluster factor for selective tree reorganization
* This system minimizes restructuring overhead by limiting updates to affected clusters rather than the entire database.

**2. Context-Preserving Centroid Calculation**

* Introduction of inverse multiplying weights during centroid calculation
* Enhanced feature preservation through:
  + Dynamic weight assignment based on similarity distances
  + Special consideration for border points
  + Iterative centroid refinement process
* This approach maintains semantic relevance at higher hierarchical levels.

**3. Skip Algorithm for Multi-Cluster Management**

* Two-phase clustering approach:
  + Initial clustering with skip mechanism for ambiguous points
  + Secondary clustering of skipped points
  + Final multi-cluster assignment for remaining ambiguous points
* This ensures appropriate handling of vectors that naturally belong to multiple clusters.

The integration of these components creates a more efficient and maintainable vector database system that addresses the core challenges of current implementations while providing improved search performance and accuracy.