Hybrid Search: BM25 & Vector Search with Filters

Hybrid Search combines multiple search paradigms, typically keyword-based (e.g., BM25) and semantic (vector) search, along with filtering capabilities to provide a comprehensive and flexible retrieval system.

Core Components

1. **BM25 (Best Match 25)**:
   - A probabilistic retrieval model that ranks documents based on the appearance of query terms in each document.
   - Considers term frequency, inverse document frequency, and document length.
2. **Vector Search**:
   - Based on dense vector representations (embeddings) of documents and queries.
   - Uses similarity measures (e.g., cosine similarity) in the vector space to find relevant documents.
3. **Filters**:
   - Additional criteria applied to narrow down the search results based on metadata or other structured information.

Process

1. **Query Processing**:
   - The input query is processed for both BM25 (tokenization, stopword removal) and vector search (embedding generation).
2. **BM25 Retrieval**:
   - The processed query is used to retrieve documents using the BM25 algorithm.
3. **Vector Search**:
   - The query embedding is used to find similar documents in the vector space.
4. **Result Combination**:
   - Results from BM25 and vector search are combined, often using a weighted approach or more sophisticated fusion techniques.
5. **Filtering**:
   - The combined results are filtered based on specified criteria (e.g., date range, document type, author).
6. **Final Ranking**:
   - The filtered results are ranked to produce the final output.

Advantages

1. **Complementary Strengths**:

- BM25 excels at precise keyword matching.
- Vector search captures semantic similarity and can handle synonyms and related concepts.

2. **Improved Relevance**: The combination often leads to more relevant results than either method alone.
3. **Flexibility**: Filters allow for fine-grained control over the result set.
4. **Handling Various Query Types**: Effective for both keyword-based and more conceptual queries.
5. **Metadata Utilization**: Filters leverage structured metadata for more targeted retrieval.

Challenges and Considerations

1. **Complexity**: Implementing and tuning a hybrid system is more complex than single-method approaches.
2. **Performance**: Running multiple retrieval methods and combining results can impact response time.
3. **Balancing Methods**: Determining the right balance between BM25 and vector search can be challenging and may vary by use case.
4. **Index Management**: Maintaining separate indexes for BM25 and vector search adds operational complexity.
5. **Filter Performance**: Poorly designed filters can significantly impact search performance, especially with large datasets.

Implementation Considerations

1. **Indexing**:
   - BM25: Inverted index for efficient keyword lookup.
   - Vector Search: Approximate Nearest Neighbor (ANN) index for fast similarity search.
2. **Weighting and Fusion**:
   - Decide how to combine BM25 and vector search results (e.g., linear combination, learning-to-rank approaches).
3. **Query Understanding**:
   - Implement logic to route queries to the most appropriate method or adjust weights based on query characteristics.
4. **Filter Optimization**:
   - Design filters for efficient execution, considering data distribution and common query patterns.
5. **Scalability**:
   - Consider distributed architectures for large-scale deployments.
6. **Relevance Tuning**:

- Implement mechanisms for continuous relevance improvement based on user feedback and metrics.

Use Cases

Hybrid Search with BM25, Vector Search, and Filters is particularly useful in:

1. **Enterprise Search**: Combining keyword matching with semantic understanding and metadata filtering.
2. **E-commerce**: Balancing exact product matches with related items and applying filters for attributes like price, category, etc.
3. **Content Management Systems**: Enabling both precise content retrieval and discovery of related materials.
4. **Legal and Medical Document Retrieval**: Where both exact term matching and conceptual similarity are crucial, along with metadata-based filtering.
5. **Research Databases**: Allowing for precise keyword search, concept-based exploration, and filtering by various academic metadata.

By leveraging the strengths of multiple search paradigms and providing filtering capabilities, Hybrid Search offers a powerful and flexible approach to information retrieval, capable of handling a wide range of search scenarios and user needs.