

# Hybrid Search: Combining BM25 and Semantic Search for Better Results with Langchain



Akash A Desai · [Follow](#)

Published in LanceDB · 4 min read · Dec 9, 2023



169



2



Have you ever wondered how search engines find exactly what you're looking for? Most often a combination of keyword matching and semantic search is used to search for user queries. This is known as hybrid search. Let's see how we can implement a simple hybrid search pipeline for document search.

## Understanding BM25:

BM25 is a ranking algorithm used in information retrieval systems to estimate the relevance of documents to a given search query.

- **What it does:** It looks at how often your search words appear in a document and considers the document's length to provide the most relevant results.

- **Why it's useful:** It's perfect for sorting through huge collections of documents, like a digital library, without bias towards longer documents or overused words.

## Key elements of BM25:

- **Term Frequency (TF):** This counts how many times your search terms appear in a document.
- **Inverse Document Frequency (IDF):** This gives more importance to rare terms, making sure common words don't dominate.
- **Document Length Normalization:** This ensures longer documents don't unfairly dominate the results.
- **Query Term Saturation:** This stops excessively repeated terms from skewing the results.

Overall

$$\text{score}(d, q) = \sum (\text{tf}(i, d) * \text{idf}(i) * (k_1 + 1)) / (\text{tf}(i, d) + k_1 * (1 - b + b * (d$$

## When is BM25/ Keyword search Ideal?

- **Large Document Collections:** Perfect for big databases where you need to sort through lots of information.
- **Preventing Bias:** Great for balancing term frequency and document length.

- **General Information Retrieval:** Useful in various search scenarios, offering a mix of simplicity and effectiveness.

## Practical Application: Building a Hybrid Search System

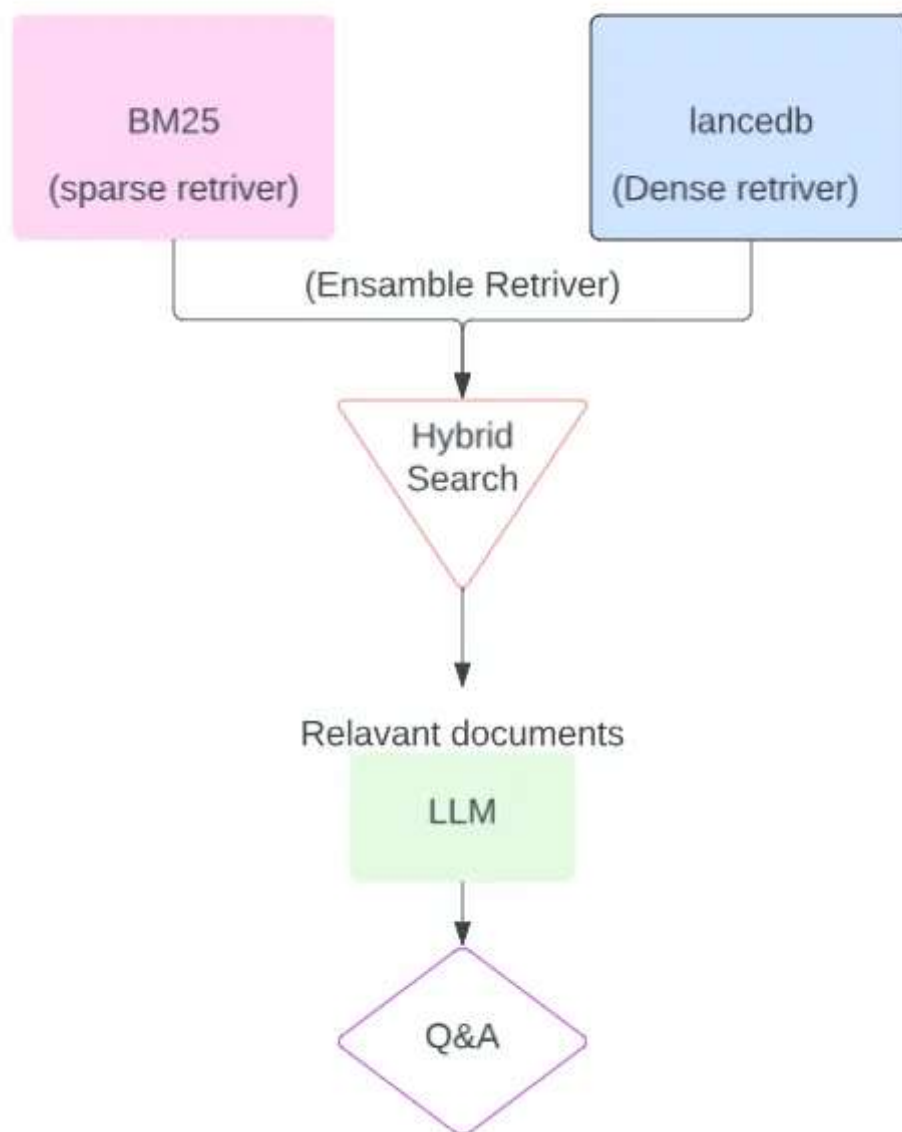
Imagine you're crafting a search system for a large digital library. You want it not only to find documents with specific keywords but also to grasp the context and semantics behind each query. Here's how:

- **Step 1:** BM25 quickly fetches documents with the search keywords.
- **Step 2:** VectorDB digs deeper to find contextually related documents.
- **Step 3:** The Ensemble Retriever runs both systems, combines their findings, and reranks the results to present a nuanced and comprehensive set of documents to the user.

## What Exactly is Hybrid Search?

Hybrid search can be imagined as a magnifying glass that doesn't just look at the surface but delves deeper. It's a two-pronged approach:

- **Keyword Search:** This is the age-old method we're most familiar with. Input a word or a phrase, and this search hones in on those exact terms or closely related ones in the database or document collection.
- **Vector Search:** Unlike its counterpart, vector search isn't content with mere words. It works using semantic meaning, aiming to discern the query's underlying context or meaning. This ensures that even if your words don't match a document exactly if the meaning is relevant, it'll be fetched.



Follow along with this colab

### Google Colaboratory

Edit description

[colab.research.google.com](https://colab.research.google.com)

Let's get to the code snippets. Here we'll use langchain with LanceDB vector store

```
# example of using bm25 & lancedb -hybrid serch

from langchain.vectorstores import LanceDB
import lancedb
from langchain.retrievers import BM25Retriever, EnsembleRetriever
from langchain.schema import Document
from langchain.embeddings.openai import OpenAIEmbeddings
from langchain.document_loaders import PyPDFLoader

# Initialize embeddings
embedding = OpenAIEmbeddings()
```

and load a single PDF.

```
# load single pdf

loader = PyPDFLoader("/content/Food_and_Nutrition.pdf")
pages = loader.load_and_split()
```

Create BM25 sparse keyword matching retriever

```
# Initialize the BM25 retriever
bm25_retriever = BM25Retriever.from_documents(pages)
bm25_retriever.k = 2 # Retrieve top 2 results
```

Create lancedb vector store for dense semantic search/retrieval.

```
db = lancedb.connect('/tmp/lancedb')
table = db.create_table("pandas_docs", data=[
    {"vector": embedding.embed_query("Hello World"), "text": "Hello World", "id":
}], mode="overwrite")

# Initialize LanceDB retriever
docsearch = LanceDB.from_documents(pages, embedding, connection=table)
retriever_lancedb = docsearch.as_retriever(search_kwargs={"k": 2})
```

Now ensemble both retrievers, here you can assign the weightage to it.

```
# Initialize the ensemble retriever
ensemble_retriever = EnsembleRetriever(retrievers=[bm25_retriever, retriever_lan
                                         weights=[0.4, 0.6])

# Example customer query
query = "which food needed for building strong bones and teeth ?
        which Vitamin & minerals importat for this?"
```

Open in app ↗

Sign up

Sign in

 Medium

 Search

 Write



Using ensemble retriever it's trying to search each word in documents, such as strong bones & teeth as well as its also searching it in using lancedb which will find most similar documents based on similarity.

```
from langchain.chat_models import ChatOpenAI

llm = ChatOpenAI(openai_api_key="sk-yourapikey")
```

```
#if you want to use opensource models such as llama,mistral check this
# https://github.com/lancedb/vectordb-recipes/blob/main/tutorials/chatbot_using_

qa = RetrievalQA.from_chain_type(llm=llm, chain_type="stuff", retriever=ensemble

query = "what nutrition needed for pregnant women  "
qa.run(query)
```

again here its searching the keyword — “ nutrition pregnant women” in the database using bm25 & returning the best matching results & similarly at the same time we are using lancedb for this. this is how it's more powerful to extract text.

```
[10] from langchain.chat_models import ChatOpenAI

llm = ChatOpenAI(openai_api_key="sk-m52pWgWbL2120p36vjLZT381bkFJB3eK9W0ue660hq3In3ZH")

qa = RetrievalQA.from_chain_type(llm=llm, chain_type="stuff", retriever=ensemble_retriever)

query = "what nutrition needed for pregnant women  "
qa.run(query)
```

'Pregnant women need to consume a balanced diet that includes a variety of foods from the main food groups. They should also have one additional meal in addition to the three meals they receive daily. Some specific nutrients that are important for pregnant women include:\n\n1. Iron: Pregnant women should consume foods rich in iron, such as lean meats, poultry, fish, beans, and leafy green vegetables. They may also need to take iron supplements as prescribed by a healthcare professional.\n\n2. Folate: Pregnant women should take folate tablets daily in addition to consuming foods rich in folate, such as dark leafy greens, citrus fruits, beans, and fortified grains.\n\n3. Calcium: Foods rich in calcium, such as milk, yogurt, cheese, and fortified plant-based milk alternatives, are important for the development of the baby's skeletal structure.\n\n4. Vitamin A: Pregnant women should consume foods rich in vitamin A, such as carrots, sweet potatoes, spinach, and mangoes. They may also need to...

```
query = "which food needed for building strong bones and teeth ? which Vitamin & minerals importat for this? "
qa.run(query)
```

'The foods that are important for building strong bones and teeth are:\n\n1. Calcium: Milk and milk products, fish with bones (such as sardines), dark green vegetables (such as broccoli and kale).\n\n2. Vitamin D: Sun exposure is the primary source, but it can also be found in vitamin D-fortified milk, eggs, fish-liver oils, and fatty fish like herring, mackerel, and salmon.\n\n3. Vitamin K: Vegetables such as spinach, lettuce, cauliflower, and cabbage, broccoli, fish, liver, meat, and eggs.\n\n4. Magnesium: Legumes, whole-grain cereals, nuts, and dark-green vegetables, as well as meat, seafood, and dairy products.\n\n5. Fluoride: Seafood, tea, coffee, soybeans, and sodium fluoride (which is often added to the water supply).\n\n6. Iron: Red meat, organ meats, whole-wheat products, shellfish, nuts, and dried fruit.\n\nIt's important to note that these nutrients work together and are best obtained through a balanced diet.'

below are answers from the traditional rag, you can check this in our repo the answers may vary based on different parameters, models, etc.

```
75 query = "what nutrition needed for pregnant women ."
qa.run(query)

'Pregnant women need a well-balanced diet that includes a variety of foods from different food groups. Some specific nutrients that are important for pregnant women include:\n\n1. Calcium: Foods rich in calcium, such as milk and mukeene (silver fish), are important for building the baby's skeletal structure.\n\n2. Iron: Pregnant women need increased amounts of iron to prevent anemia. They can take iron supplements to meet this requirement.\n\n3. Folic Acid: Taking the required amounts of folic acid supplements is important to prevent birth defects and promote the healthy development of the baby.\n\n4. Protein: Pregnant women need extra protein for the growth and development of the baby. Good sources of protein include lean meats, poultry, fish, eggs, dairy products, and legumes.\n\n5. Vitamins and Minerals: Pregnant women have higher nutrient needs in general, so it is important to consume a variety of fruits and vegetables to get essential vitamins and minerals. Foods rich in vitamin A and iron are particularly important.\n\n6. Snacks: Pregnant women should have snacks in between meals to help meet their increased nutrient needs.\n\nIt is also important for pregnant women to maintain a healthy weight and stay hydrated by drinking plenty of water. Consulting with a healthcare professional or a registered dietitian can provide personalized guidance on nutrition during pregnancy.'
```

```
69 [29] query = "which food needed for building strong bones and teeth ? which Vitamin & minerals importat for this? "
qa.run(query)

'To build strong bones and teeth, it is important to consume foods that are rich in calcium, vitamin D, and vitamin C. \n\n1. Calcium: The major source of calcium is milk and milk products. Other sources include fish eaten with bones (such as silver fish and haplochromis species) and dark green vegetables.\n\n2. Vitamin D: This vitamin is produced by the skin on exposure to sunlight. Food sources of vitamin D include milk, cheese, butter, blue band margarine, fatty fish, eggs, and liver.\n\n3. Vitamin C: While not directly involved in bone formation, vitamin C plays a role in collagen synthesis, which is important for maintaining the structure of bones and teeth. Good sources of vitamin C include citrus fruits, berries, and leafy green vegetables.\n\nAdditionally, other minerals such as phosphorus and chromium also play a role in maintaining strong bones and teeth. Phosphorus can be obtained from sources like fish with edible bones and tofu, while chromium can be found in whole wheat and other whole grains.'
```

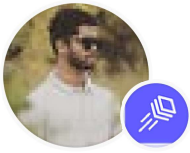
you can try this on colab with your pdf & use case. This is how you can use hybrid search to improve your search quality.

## Explore More with Our Resources

Discover the full potential of hybrid search and beyond in our LanceDB repository, offering a setup-free, persisted vectorDB that scales on on-disk storage. For a deep dive into applied GenAI and vectorDB applications, examples, and tutorials, don't miss our VectorDB-Recipes at <https://github.com/lancedb/vectordb-recipes>. From advanced RAG methods like Flare, Rerank, and HyDE to practical use cases, our resources are designed to inspire your next project .

[Llm](#)[Vector Database](#)[Semantic Search](#)





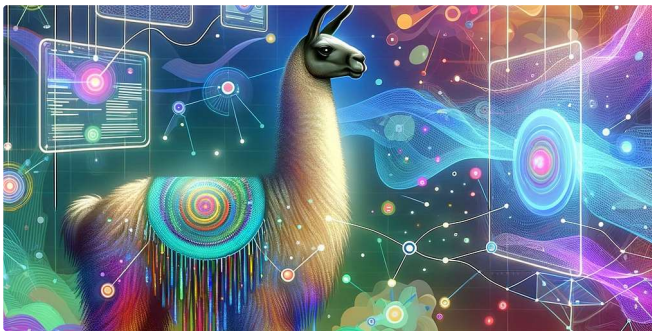
## Written by Akash A Desai

Follow

77 Followers · Writer for LanceDB

Data Scientist with 4+ years of exp ! open source ! Vision ! Generative AI ! Vectordb ! llms

### More from Akash A Desai and LanceDB



Akash A Desai in LanceDB

### Better RAG with Active Retrieval Augmented Generation FLARE

Welcome to our deep dive into Forward-Looking Active Retrieval Augmented...

7 min read · Nov 18, 2023



195



202



2

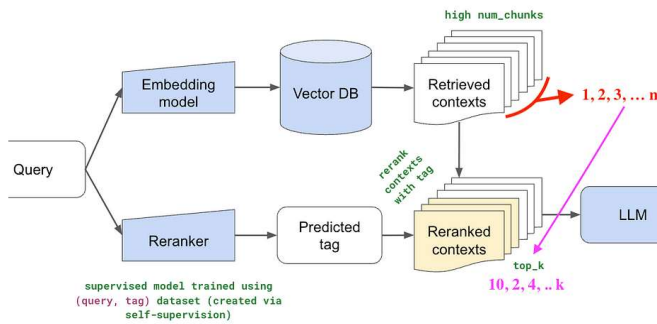


Chang She in LanceDB

### LanceDB + Polars

A (near) perfect match

6 min read · Jan 19, 2024



Mahesh Deshwal in LanceDB

## Simplest Method to improve RAG pipeline: Re-Ranking

Problem Statement:

4 min read · Nov 23, 2023



176



Akash A Desai in LanceDB

## Better RAG with LOTR-Lord of Retriever

Implementing LOTR merger retriever & solving the 'Lost in the Middle' Challenge in...

7 min read · Dec 21, 2023



121



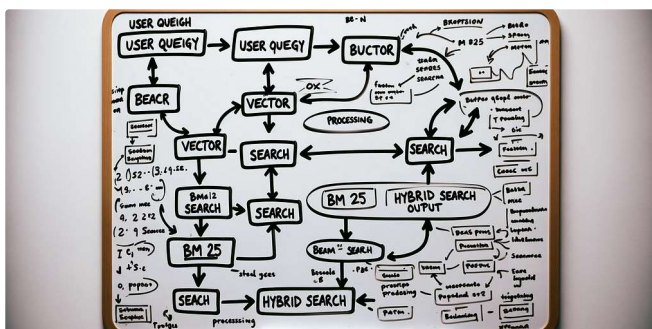
1



See all from Akash A Desai

See all from LanceDB

## Recommended from Medium





azhar in azhar labs

## Vector Search and BM 25 (hybrid search): Advanced Ensemble...

Introduction

🌟 · 5 min read · Oct 14, 2023



77



Fanghua (Joshua) Yu in Neo4j Developer Blog

## Building A Graph+LLM Powered RAG Application from PDF...

A Step-by-step Walkthrough with GenAI-Stack and OpenAI

6 min read · Jan 15, 2024



544



3



### Lists



#### Natural Language Processing

1215 stories · 692 saves



#### data science and AI

40 stories · 80 saves



#### ChatGPT prompts

42 stories · 1146 saves



#### Generative AI Recommended Reading

52 stories · 743 saves

Hybrid Search...



Leonie Monigatti in Towards Data Science

## Improving Retrieval Performance in RAG Pipelines with Hybrid Search

How to find more relevant search results by combining traditional keyword-based search...

8 min read · Nov 28, 2023

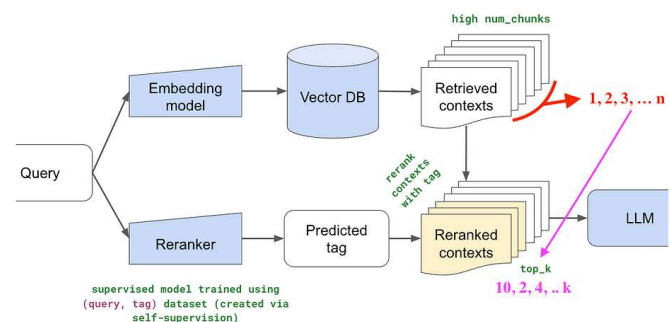


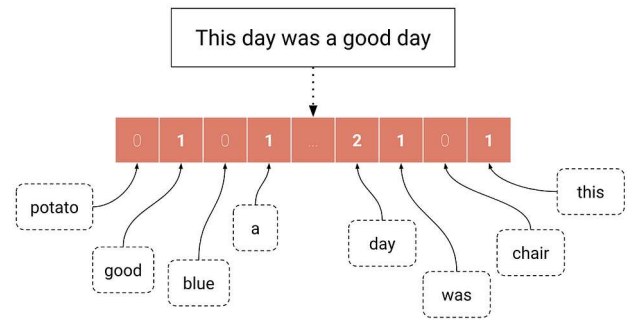
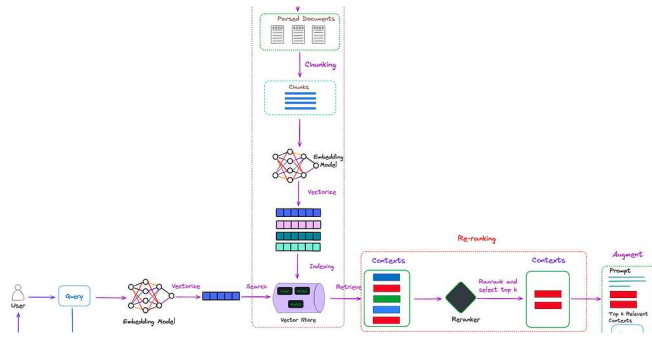
Mahesh Deshwal in LanceDB

## Simplest Method to improve RAG pipeline: Re-Ranking

Problem Statement:

4 min read · Nov 23, 2023





Florian June in Towards AI

## Advanced RAG 04: Re-ranking

From Principles to Two Mainstream Implementation Methods

★ • 11 min read • Feb 14, 2024



Anar Aliyev in LLMRails

## Write your own Hybrid Search with RRF and 'Ember' rerank

Embedding models are good for vector search with the help of vector databases. Bu...

3 min read • Oct 23, 2023

See more recommendations