Vector database and its applications

Vladimir Egay · Follow
6 min read · Jun 16

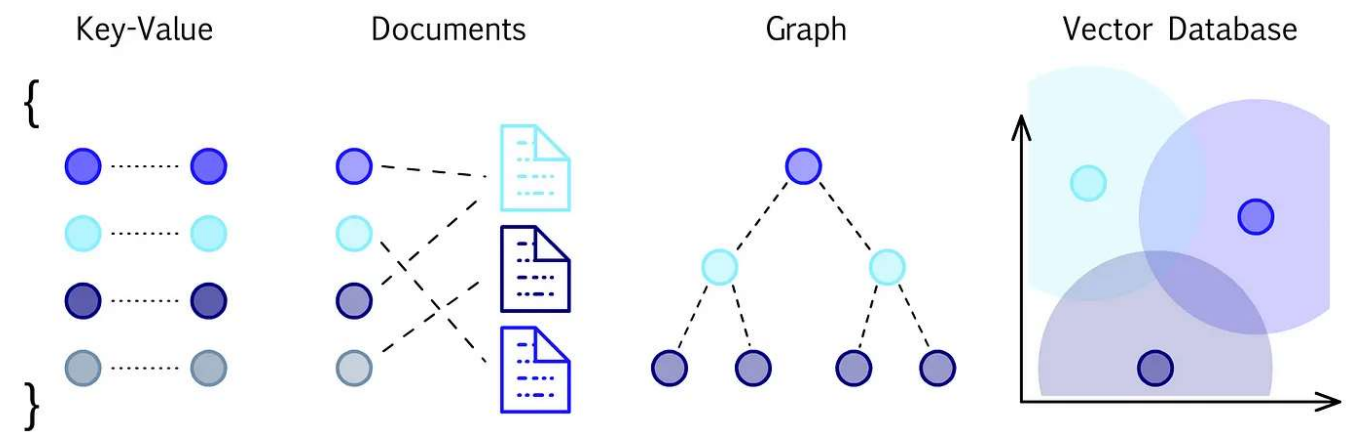

Image source: pinecone.io

I. Introduction

In the modern landscape of big data and artificial intelligence, the way we store and retrieve data has significantly evolved. With the rise of complex, multi-dimensional data, innovative management solutions are more crucial than ever. One such groundbreaking solution is Vector Databases. These databases, unlike traditional ones, don't rely on structured tabular data. Instead, they efficiently handle high-dimensional vector data, making them particularly useful for data types like images, audio, video, and text, where semantic similarities can be encapsulated in vector forms.

Vector databases, also known as similarity search engines or approximate nearest neighbor (ANN) search engines, revolutionize how we perceive data retrieval in high-dimensional spaces. They allow for the identification of data that is "closest" to a given vector even within extremely large datasets. Their applications are widespread, from powering advanced recommendation systems and semantic searches to driving sophisticated AI models. In this blog post, we'll dive deeper into vector databases, their benefits over traditional databases, their applications, real-world examples, and the promising future they hold in the realm of big data and AI.

II. Deep Dive into Vector Databases

Vector databases, also known as similarity or nearest neighbor search databases, operate on a simple yet powerful premise: They allow us to search for data points in a high-dimensional space close to a given query point, also known as ‘nearest neighbors’. Rather than operating on a strict table-like structure like traditional databases, vector databases use geometrical positions in a multi-dimensional space.

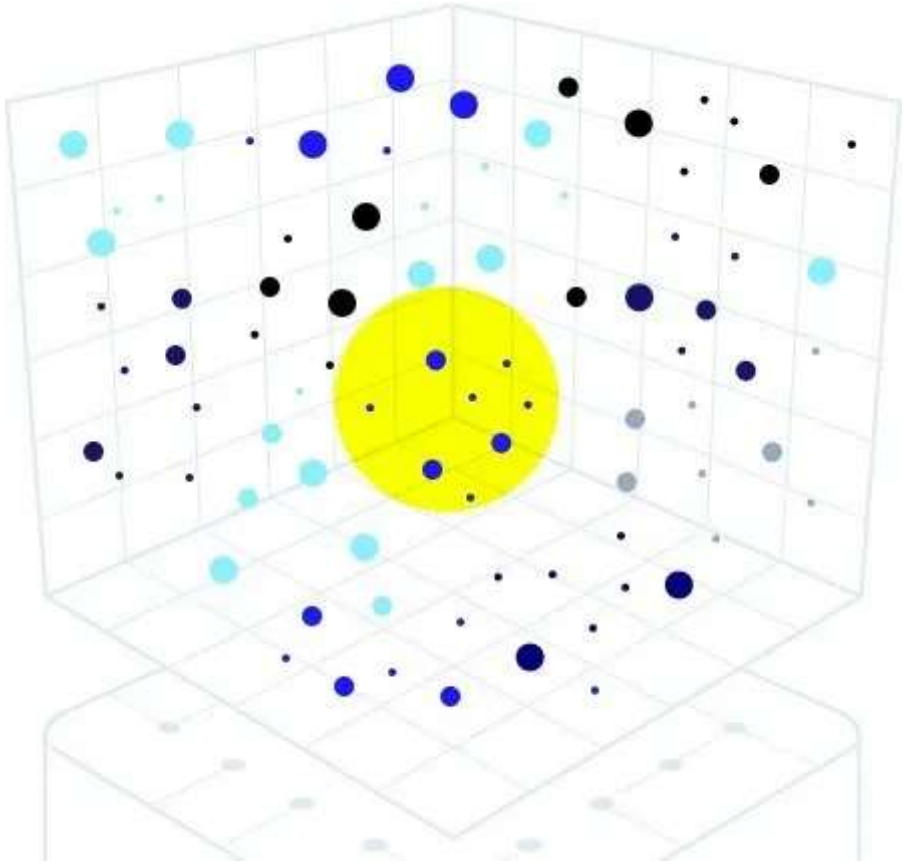

Image source: pinecone.io

A key characteristic of vector databases is their ability to handle unstructured data. While traditional databases excel at handling structured data, they fall short when it comes to complex and unstructured data types. Vector databases excel in this area, as they can efficiently store and search through vast amounts of unstructured data. Additionally, they employ distance metrics to quantify the 'similarity' between vectors, providing a quantitative measure of how closely related two pieces of data are.

There are various types of vector databases, each with its own set of algorithms and methods to efficiently store and retrieve data. For instance, exact search databases retrieve the most accurate results by checking the distance from the query vector to every other vector in the database. However, this can be computationally intensive for large datasets. On the other hand, approximate nearest neighbor search databases aim to balance speed and accuracy. They might not always return the most accurate results, but they do so much more quickly and with fewer computational resources, making them suitable for large-scale, real-time applications.

Examples of these databases might include:

- Pinecone, a fully managed vector database
- Weaviate, an open-source vector search engine

- Redis as a vector database
- Milvus, a vector database built for scalable similarity search
- Chroma, an open-source embeddings store
- Zilliz, data infrastructure, powered by Milvus

IV. Applications of Vector Databases

There are multiple ways you can use Vector Database and there are most known ones.

1. Search Engines: Vector databases play an integral part in improving the accuracy and efficiency of search engines. By converting search queries and web content into vectors, search engines can quickly find the most relevant results. Search query could be anything that can be represented as vectors from text to image and audio.

2. Recommendation Systems: Online platforms like Netflix, Amazon, and Spotify utilize vector databases to improve their recommendation systems. By transforming user-profiles and item descriptions into vectors, these systems can find similarities and make accurate recommendations that align with the user's interests.

3. Natural Language Processing and Understanding: In NLP, vector databases are utilized to understand and process human language more efficiently. Text data is transformed into word vectors, allowing AI models to understand the semantic and syntactic similarities between different words and phrases.

4. Image and Video Recognition: Vector databases assist in improving the accuracy of image and video recognition software. By converting images and videos into vectors, these systems can quickly identify patterns, objects, or faces. This application is widely used in fields like surveillance, social media, and autonomous vehicles.

5. Genomics and Bioinformatics: In genomics, vector databases help in searching and matching genetic sequences. It allows for quick and precise comparison of large volumes of genetic data, aiding in tasks like disease prediction, genetic modifications, and drug discovery.

V. Real-World Application of Vector Database: A Personal Experience with Milvus

1. Understanding the Challenge

Our project involved developing an efficient method for users of a Q&A website to identify if their questions had already been fielded and addressed. We observed that traditional techniques such as tag or word similarity search were not sufficient due to their inability to capture the semantics of queries. For instance, a word similarity search for an “Apple product” (referring to the tech company) might return results about apples the fruit, resulting in irrelevant matches. Moreover, our solution needed to withstand high traffic volumes and cater to a multilingual user base.

2. Solution Overview:

To address these challenges, we turned towards text embedding solutions, with OpenAI’s LLM-based text embedding being our choice. It appealed to us because of its easy-to-use APIs, high-quality embeddings (thanks to training on extensive datasets), and cost-effectiveness.

3. Detailing the Implementation Process:

A. **Converting Questions into Vectors:** Our process initiates when a user enters their question in the search bar. We use OpenAI’s embedding model to transform the question into a vector, effectively capturing its semantic meaning.

B. **Comparing and Ranking:** This newly formed vector is then compared to our database of pre-existing questions. The system sorts all the questions based on their similarity scores, a feature that vector databases usually provide.

C. **Threshold-Based Suggestions:** We have set a predefined threshold for similarity scores. If all the scores are below this threshold, the system prompts the user to ask their question. However, if relevant matches are found, the system lists all similar questions along with their answers that it discovered in the database.

4. Future development:

Moving forward, we are considering expanding the system's capabilities by integrating it with ChatGPT. This upgrade would allow the AI to produce personalized responses using the top 'k' similar questions and their answers as context. Essentially, we'll be instructing ChatGPT to distill the relevant information from the provided context and generate a unique answer to the user's query.

This step is a natural progression, given our existing process for extracting the 'most relevant content.' Not only will this improve the user experience,

but it also showcases a valuable use case for AI models like ChatGPT in leveraging corporate data to deliver personalized and efficient customer interactions.

In conclusion, this reflection on potential future developments underscores the pivotal role vector databases can play in enhancing and optimizing content management in various scenarios, especially those involving large volumes of multilingual data.

VI. The Future of Vector Databases

In the present era marked by the ascendancy of Language Learning Models (LLMs) and the increasing prominence of Artificial Intelligence, vector databases are witnessing an unparalleled surge in popularity. This heightened interest is not limited to well-established tech giants but also extends to smaller startups, which are attracting substantial investments.

A prime exemplar is Pinecone, a company that recently secured a staggering investment of $100 million, propelling its valuation to $750 million in April of this year. [Source] Remarkably, this financial achievement occurred amidst a generally conservative investment climate. Pinecone's journey began just in 2021, on the cusp of the AI explosion, and its rapid growth is a

testament to the strategic significance of vector databases. Given that LLMs require long-term memory capabilities, the relevance and popularity of vector databases are poised to continue their upward trajectory for the foreseeable future.

VII. Conclusion

In conclusion, the advent of vector databases is significantly transforming the management of multi-dimensional data in the era of big data and AI. Excelling at handling unstructured data like images, audio, video, and text, these databases are able to quantify ‘similarity’ in high-dimensional spaces. They play pivotal roles in applications such as recommendation systems, semantic search engines, NLP models, and image and video recognition software. A real-world case study with a Q&A website illustrates their effectiveness in managing large volumes of data, particularly in high-traffic and multilingual contexts.

Looking ahead, the rising popularity of AI and the advent of Language Learning Models (LLMs) suggest a promising future for vector databases. Their relevance has been recognized even in a conservative investment environment, as evidenced by substantial investments in startups like

Pinecone. As the complexities of data continue to grow, vector databases are set to remain a vital tool in the data management landscape.

Vector Database | Text Embedding | OpenAI | Semantic Search