Google’s New MUVERA Algorithm Improves Search via @sejournal, @martinibuster

Google announced a new multi-vector retrieval algorithm called MUVERA that speeds up retrieval and ranking, and improves accuracy. The algorithm can be used for search, recommender systems (like YouTube), and for natural language processing (NLP).

Although the announcement did not explicitly say that it is being used in search, the research paper makes it clear that MUVERA enables efficient multi-vector retrieval at web scale, particularly by making it compatible with existing infrastructure (via MIPS) and reducing latency and memory footprint.

Vector Embedding In Search

Vector embedding is a multidimensional representation of the relationships between words, topics and phrases. It enables machines to understand similarity through patterns such as words that appear within the same context or phrases that mean the same things. Words and phrases that are related occupy spaces that are closer to each other.

- The words “King Lear” will be close to the phrase “Shakespeare tragedy.”
- The words “A Midsummer Night’s Dream” will occupy a space close to “Shakespeare comedy.”
- Both “King Lear” and “A Midsummer Night’s Dream” will be located in a space close to Shakespeare.

The distances between words, phrases and concepts (technically a mathematical similarity measure) define how closely related each one is to the other. These patterns enable a machine to infer similarities between them.

MUVERA Solves Inherent Problem Of Multi-Vector Embeddings

The MUVERA research paper states that neural embeddings have been a feature of information retrieval for ten years and cites the ColBERT multi-vector model research paper from 2020 as a breakthrough but that says that it suffers from a bottleneck that makes it less than ideal.

> *"Recently, beginning with the landmark ColBERT paper, multi-vector models, which produce a set of embedding per data point, have achieved markedly superior performance for IR tasks. Unfortunately, using these models for IR is computationally expensive due to the increased complexity of multi-vector retrieval and scoring."*

Google's announcement of MUVERA echoes those downsides:

> *"… recent advances, particularly the introduction of multi-vector models like ColBERT, have demonstrated significantly improved performance in IR tasks. While this multi-vector approach boosts accuracy and enables retrieving more relevant documents, it introduces substantial computational challenges. In particular, the increased number of embeddings and the complexity of multi-vector similarity scoring make retrieval significantly more expensive."*

Could Be A Successor To Google's

RankEmbed Technology?

The United States Department of Justice (DOJ) antitrust lawsuit resulted in testimony that revealed that one of the signals used to create the search engine results pages (SERPs) is called RankEmbed, which was described like this:

> *"RankEmbed is a dual encoder model that embeds both query and document into embedding space. Embedding space considers semantic properties of query and document in addition to other signals. Retrieval and ranking are then a dot product (distance measure in the embedding space)… Extremely fast; high quality on common queries but can perform poorly for tail queries…"*

MUVERA is a technical advancement that addresses the performance and scaling limitations of multi-vector systems, which themselves are a step beyond dual-encoder models (like RankEmbed), providing greater semantic depth and handling of tail query performance.

The breakthrough is a technique called Fixed Dimensional Encoding (FDE), which divides the embedding space into sections and combines the vectors that fall into each section to create a single, fixed-length vector, making it faster to search than comparing multiple vectors. This allows multi-vector models to be used efficiently at scale, improving retrieval speed without sacrificing the accuracy that comes from richer semantic representation.

According to the announcement:

> *"Unlike single-vector embeddings, multi-vector models represent each data point with a set of embeddings, and leverage more sophisticated similarity functions that can capture richer relationships between datapoints.*
>
> *While this multi-vector approach boosts accuracy and enables*

> *retrieving more relevant documents, it introduces substantial computational challenges. In particular, the increased number of embeddings and the complexity of multi-vector similarity scoring make retrieval significantly more expensive.*
>
> *In 'MUVERA: Multi-Vector Retrieval via Fixed Dimensional Encodings', we introduce a novel multi-vector retrieval algorithm designed to bridge the efficiency gap between single- and multi-vector retrieval.*
>
> *…This new approach allows us to leverage the highly-optimized MIPS algorithms to retrieve an initial set of candidates that can then be re-ranked with the exact multi-vector similarity, thereby enabling efficient multi-vector retrieval without sacrificing accuracy."*

Multi-vector models can provide more accurate answers than dual-encoder models but this accuracy comes at the cost of intensive compute demands. MUVERA solves the complexity issues of multi-vector models, thereby creating a way to achieve greater accuracy of multi-vector approaches without the the high computing demands.

What Does This Mean For SEO?

MUVERA shows how modern search ranking increasingly depends on similarity judgments rather than old-fashioned keyword signals that SEO tools and SEOs are often focused on. SEOs and publishers may wish to shift their attention from exact phrase matching toward aligning with the overall context and intent of the query. For example, when someone searches for "corduroy jackets men's medium," a system using MUVERA-like retrieval is more likely to rank pages that actually offer those products, not pages that simply mention "corduroy jackets" and include the word "medium" in an attempt to match the query.

Read Google's announcement:

MUVERA: Making multi-vector retrieval as fast as single-vector search

Featured Image by Shutterstock/bluestork

https://www.searchenginejournal.com/googles-new-muvera-algorithm-improves-search/550070/