Response to “Word embeddings / vector search” on Wikitech-l - May 2023

Context:

https://lists.wikimedia.org/hyperkitty/list/wikitech-l@lists.wikimedia.org/thread/PJGRN2H2CUSACZH4QSI2VOV3WBDZ3J6O/

Proposed reply

Hello all!

I’m Guillaume, the engineering manager for the Search Platform team.

Thanks for your interest in Search! It’s always good to see that people care about it!

The search on Wikipedias is currently not using word embeddings or vector search. This is something we’ve been thinking about for a long time, so there is definitely interest on our side! A few things have been keeping us from introducing some kind of vector search so far:

We’d like to focus on the improvements to user experience more than on the tools / technologies used to achieve it. Vector search is a great tool, but what are we trying to improve with it? There might be other ways to achieve the same result that are not vector search. For example, we've worked on query suggestion / correction, which can help refine queries and provide additional results.

Currently, search is backed by Elasticsearch, which only recently started supporting vector search. As Isaac already mentioned, our current search seems to work reasonably well for the use cases we see (like 80% of auto-complete searches resulting in a success). There are still plenty of potential improvements, but migrating to a different backend just so that we could use vector search does not seem like a good use of our time. This can be revisited now that Elasticsearch provides that support. (On a side note, we will need to move away from Elasticsearch due to licensing issues. [1] The obvious alternative is OpenSearch, [2] which also supports vector search, but sadly seems to be more oriented towards log management than full-text search - a whole other story).

While vector search is really powerful, it is not entirely clear that this is what is most needed to improve the search experience at the moment. For example, the Special:MediaSearch UI changes [3] are a clear improvement to search on Commons that finally makes searching for images a reasonable experience.

The Search Platform team is also a fairly small team of 7±2 people, working on both Wikidata Query Service and Search. Just operating the search infrastructure and keeping up with the amount of data we need to ingest and store is major work and sadly does not allow us to spend as much time as we'd like to invest in new features / technologies.

The Foundation in general is a comparatively small organization for such a high-profile website and mission with such broad scope. The Search Platform team in particular is only 7±2 people, working on both the Wikidata Query Service and Search. Just operating the search infrastructure and keeping up with the amount of data we need to ingest and store is major work and sadly does not allow us to spend as much time as we'd like to invest in new features and technologies. Waiting for some features and technologies to become mainstream and easy to incorporate—in this case after our migration away from Elasticsearch—is often a better approach.

All that being said, the line is being blurred these days between Search, Machine Learning, and Natural Language Processing. There are ongoing experiments around using language models for search, which is going into the same direction as word embeddings / vector search.

Have fun!

Guillaume

[1] https://phabricator.wikimedia.org/T272111
[2] https://opensearch.org/
[3] https://commons.wikimedia.org/w/index.php?search=cat&title=Special:MediaSearch&go=Go&type=image