Orion: An interactive information retrieval system for scientific knowledge discovery

Kostas Stathoulopoulos[1], Zac Ioannidis, Lilia Villafuerte
[1]Mozilla, United Kingdom

Abstract

The volume of scientific publications has rapidly increased in recent years, partly fuelled by the widespread adoption of preprints. This newly created and often time-critical knowledge is spread across databases or stored in large, cross-disciplinary collections that are difficult to navigate with traditional, keyword-based search engines. There is a need for tools that combine academic knowledge and metaknowledge from different sources and enable the exploration of large-scale information systems by bringing human intelligence and attention more actively into the search process.

In this work, we describe Orion[1], an interactive information retrieval (IIR) [1] system for scientific texts. In Orion, we query Microsoft Academic Graph (MAG) with a journal, conference or field of study to create a domain-specific database. We enrich it by geocoding institutional affiliations and inferring the authors' gender while we also produce country-level indicators of scientific progress such as research interdisciplinarity and gender diversity. Then, we use a sentence-level DistilBERT [2] to encode paper abstracts and find their vector representation.

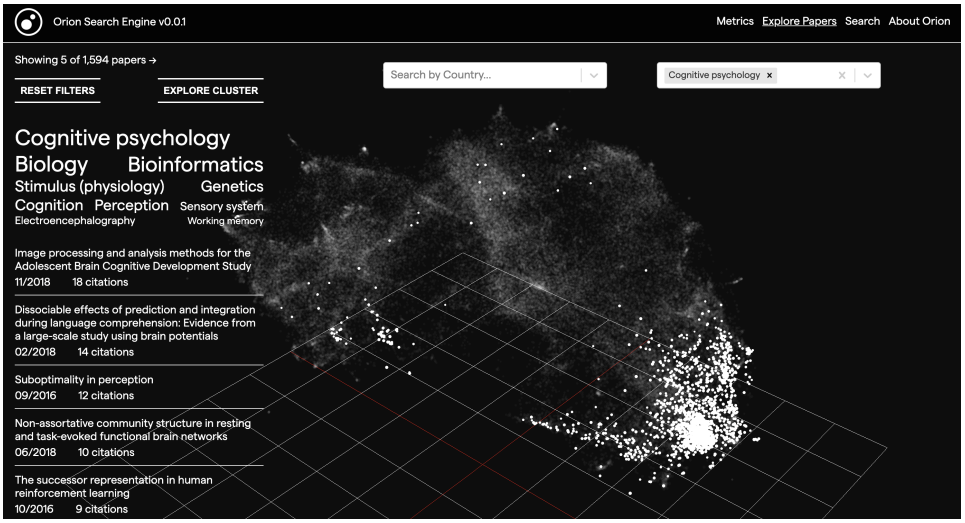

Figure 1: 3D UMAP projection of the sentence-BERT abstract embeddings. The most used Fields of Study are shown on the top left and the papers tagged as *Cognitive psychology* are highlighted in the plot.

The IIR has two main components; a vector similarity search engine that can be used for long text queries and interactive data visualisations that enable users to explore the database and refine their queries. For the former component, we build a FAISS index [3] with the abstract embeddings. We encode new queries with the sentence-BERT model and search for similar vectors in the FAISS index. For the latter, we reduce the dimensionality of the abstract vectors using UMAP [4] and visualise their 3-dimensional projection with a scatterplot where every particle is a paper and the distance between them indicates their cosine distance 1; the closer two particles are, the more semantically similar. Users can filter the search space by topic, country or by using the research indicators we developed. Lastly, they can select a subset of the particles and retrieve those papers and their metadata.

To conclude, Orion creates domain-specific and information-rich scientific databases that can be explored using an interactive information retrieval system. We believe the combination of advanced text representation methods with data visualisation improves knowledge discovery as it enables users query the database with variable-length text and iteratively narrow the search space.

[1]Orion is open-source and the alpha version uses publications from bioRxiv `https://www.orion-search.org/`

References

[1] Gary Marchionini. Toward human-computer information retrieval. *Bulletin of the American Society for Information Science and Technology*, 32(5):20–22, 2006.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] J Johnson, M Douze, and H Jégou. Billion-scale similarity search with gpus. arxiv 2017. *arXiv preprint arXiv:1702.08734*.

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.