PaperMatch: Cross-Domain Semantic Search Accelerating Scientific Discovery

Mitanshu Sukhwani
Ahmedabad, Gujarat, India.
mitanshu.sukhwani@gmail.com

Abstract

As interdisciplinary research gains prominence, tools enabling cross-domain exploration have become indispensable. PaperMatch is an innovative platform featuring three specialized websites: PaperMatch, PaperMatchBio, and PaperMatchMed. These sites focus on distinct domains, indexing abstracts from arXiv, bioRxiv, and medRxiv, respectively. Offering a semantically driven search powered by embedding-based techniques and Milvus vector storage, PaperMatch helps users find conceptually related papers across repositories. This fosters interdisciplinary insights and supports groundbreaking discoveries across a wide range of scientific disciplines. Code available on GitHub at mitanshu7/PaperMatch.

1 Introduction

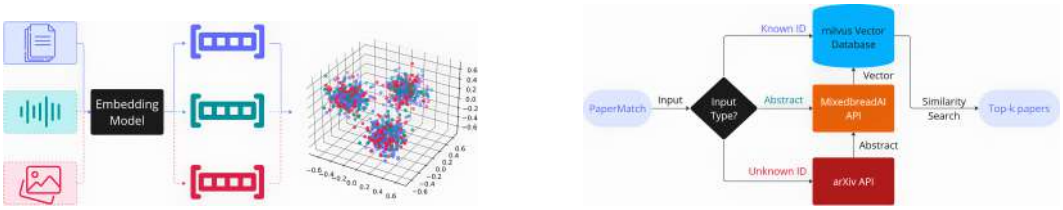
The surge in scientific publications presents significant challenges for researchers aiming to identify related work across disciplines. As of November 2024, arXiv hosts 2.6 million articles, with 24 thousand papers added in October alone. Traditional keyword-based search tools often fail to capture conceptually similar work expressed in varying terminologies. PaperMatch overcomes this limitation by leveraging semantic embeddings for cross-domain search, empowering researchers to uncover connections across arXiv, bioRxiv, medRxiv, and beyond.

2 Methodology

Each PaperMatch site fetches abstracts from its designated repository and processes these texts using the embedding model mxbai-embed-large-v1, Lee et al. [2024], accessed via the Hugging Face Transformers library, Wolf et al. [2020]. The embedding model generates vectors of length 1024 that capture the semantic content of the abstracts, max. 512 tokens, going beyond simple keywords to encapsulate complex concepts. This approach allows the model to encode similar ideas into vectors positioned closely in the embedding space, enabling accurate and contextually relevant search results. This model was selected for its high performance on Massive Text Embedding Benchmark (MTEB), Muennighoff et al. [2022], during the inception of PaperMatch as well as it's support for vector quantization, Gray [1984], and Matryoshka Representation Learning, Kusupati et al. [2024].

To speed up the embedding process, we utilized PyTorch, Paszke et al. [2019], as the computational backend on an RTX 4050 Laptop GPU. PyTorch's hardware acceleration enabled efficient handling of large datasets, scaling the embedding process for a growing corpus of scientific literature. The resulting `float32` vectors are stored in Milvus, Wang et al. [2021], a high-performance vector database optimized for fast similarity searches using nearest-neighbor algorithms. Users can search by inputting an ID or full abstracts to gather conceptually similar work. The user interface, built with Gradio, Abid et al. [2019], delivers a responsive, real-time platform for interactive exploration.

1st Conference on Applied AI and Scientific Machine Learning (CASML 2024).

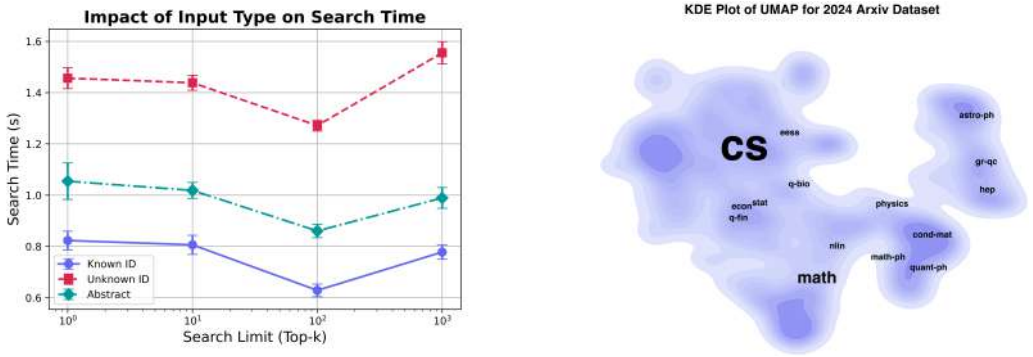


(a) (b)

Figure 1: Illustrations of the core components of the PaperMatch system. (a) Embedding Models generate high-dimensional fixed-length vectors, (b) Flowchart for PaperMatch.

3 Results and Discussion

Preliminary user feedback underscores PaperMatch's potential to advance cross-disciplinary research by surfacing related works that might otherwise be overlooked in traditional keyword-based searches, fostering collaboration and innovation. As shown in Figure 2a, Milvus demonstrates exceptional efficiency, consistently computing cosine similarity for 2.6 M records in well under one second on VM.Standard.A1.Flex (4 CPU, 24 GB Memory). To explore the current scientific landscape, we employ Uniform Manifold Approximation and Projection (UMAP), McInnes et al. [2018], to reduce high-dimensional embeddings to 2D. The resulting scatter plot is further refined using Kernel Density Estimation (KDE), Parzen [1962], to highlight structural patterns, as illustrated in Figure2b.



(a) (b)

Figure 2: Visualizing the capabilities and insights provided by PaperMatch. (a) Performance of PaperMatch, (b) Map of arXiv for the year 2024.

4 Conclusion

The PaperMatch platform demonstrates the power of embedding-based search for enabling cross-domain knowledge discovery. By integrating state-of-the-art embedding techniques and vector databases, PaperMatch provides a valuable tool for the scientific community, offering a novel approach to literature search that prioritizes semantic relevance. This platform holds significant potential for supporting interdisciplinary innovation and fostering connections between researchers across scientific disciplines. However, PaperMatch inherits biases, interpretability issues, and domain-specific performance gaps, Rakivnenko et al. [2024], from the embedding models it relies on, while vector databases add complexity by treating embeddings as independent data points, Arye and Sewrathan [2024]. Additionally, the current user interface, though functional, would benefit from enhanced clarity, navigation, and personalization features to maximize user engagement. Future developments will focus on addressing these limitations, expanding the platform to support additional repositories, improving the embedding model's capacity for nuanced contextual understanding, and incorporating user feedback to further enhance search accuracy and the overall user experience.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. URL https://arxiv.org/abs/1906.02569.

Matvey Arye and Avthar Sewrathan. Vector databases are the wrong abstraction, 2024. URL https://www.timescale.com/blog/vector-databases-are-the-wrong-abstraction/.

Robert M. Gray. Vector quantization. *IEEE ASSP Magazine*, 1(2):4–29, 1984. doi: 10.1109/MASSP.1984.1162229. URL https://doi.org/10.1109/MASSP.1984.1162229.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. 2024. URL https://arxiv.org/abs/2205.13147.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embedding model, 2024. URL https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. URL https://arxiv.org/abs/1802.03426.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962. doi: 10.1214/aoms/1177704472. URL https://doi.org/10.1214/aoms/1177704472.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. 2019. URL https://arxiv.org/abs/1912.01703.

Vasyl Rakivnenko, Nestor Maslej, Jessica Cervi, and Volodymyr Zhukov. Bias in text embedding models. 2024. URL https://arxiv.org/abs/2406.12138.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627, 2021. URL https://doi.org/10.1145/3448016.3457550.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. 2020. URL https://arxiv.org/abs/1910.03771.