Proxima: Enhanced Search Engine for Research Articles

Ceydanur Şen - Nurullah Uçan

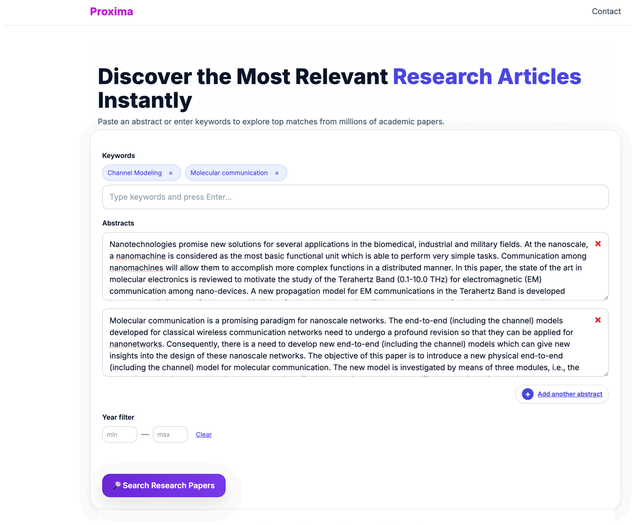
Boğaziçi University - CMPE 492 Project - Advisor: H. Birkan Yılmaz

Motivation and Problem Definition

- Finding relevant academic papers is time-consuming and often overwhelming for researchers.
- Traditional keyword-based search engines frequently return many irrelevant or low-quality results.
- Existing academic search tools do not fully exploit modern semantic search and rich metadata.
- Researchers need a system that surfaces recent, high-quality papers tailored to their specific interests.

Objectives and Contributions

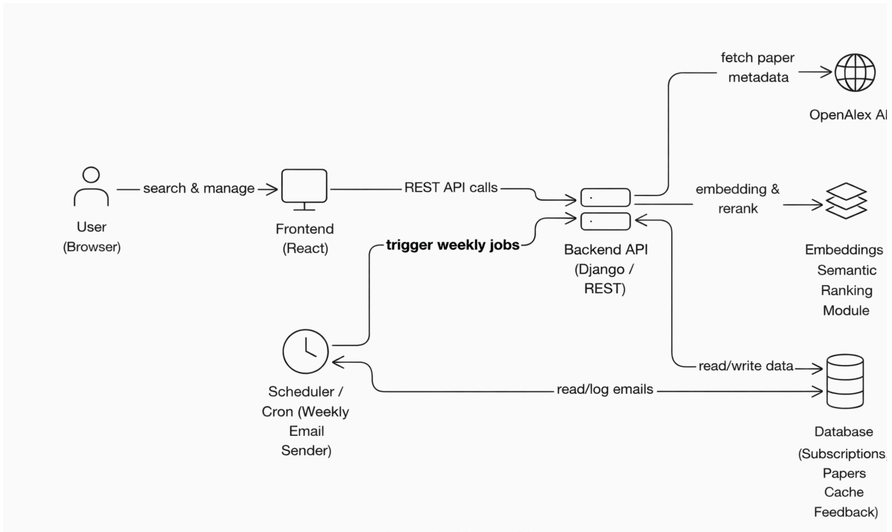
- Design and implement an enhanced academic search engine on top of the OpenAlex dataset.
- Combine multi-abstract input, semantic embeddings, and keyword-based retrieval to improve result relevance.
- Provide a web interface for exploring, saving, and managing topic-focused **paper recommendations.**
- Support long-term topic tracking through **email subscriptions** and user feedback collection.



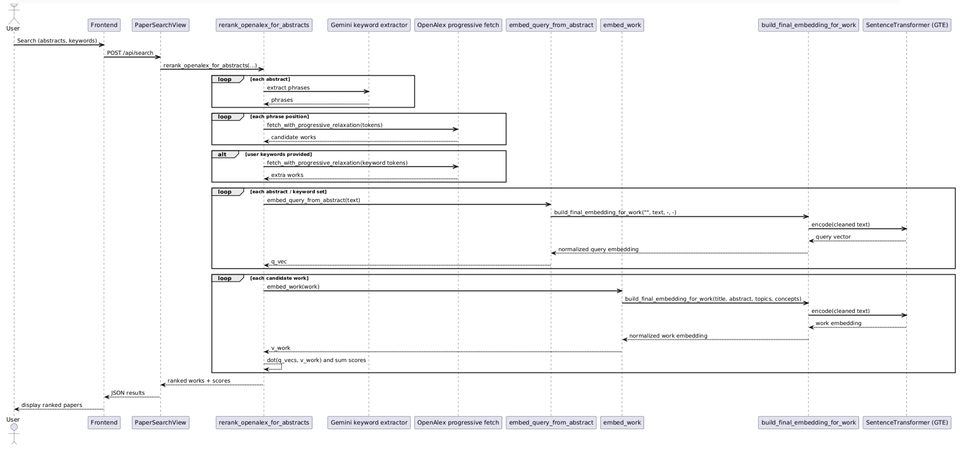
Home Page View of Proxima

Methodology

- The user submits keywords and/or one or more seed abstracts, optionally with a publication year range.
- The backend retrieves candidate papers from OpenAlex using keyword-based filtering and metadata constraints.
- We encode seed abstracts and candidate papers with a pretrained sentence embedding model (Alibaba-NLP/gte-large-en-v1.5 via Sentence-Transformers, 1024-D), and re-rank OpenAlex candidates by semantic similarity.
- The top-ranked papers are returned to the frontend, where users can inspect results, create subscriptions, and provide feedback.



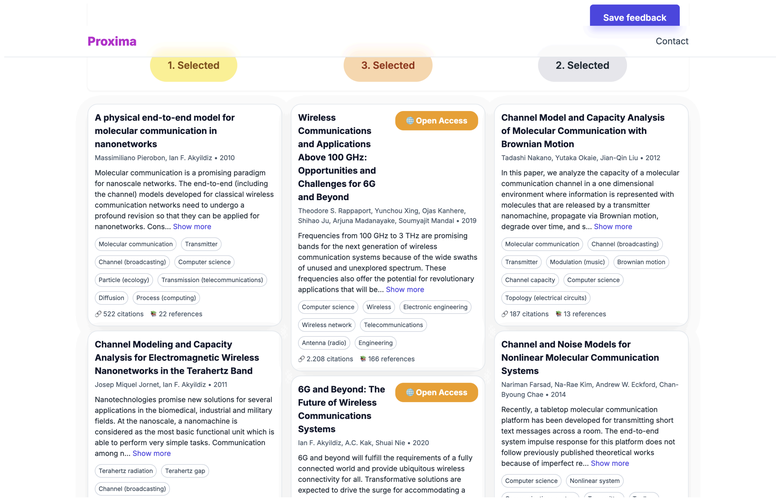
System Architecture Diagram



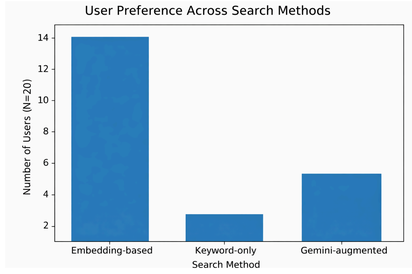
Sequence Diagram of Embedding-based multi-abstract search pipeline

Experiments and Results

- We compare three search methods: embedding-based search, raw OpenAlex keyword-only search, and Gemini-augmented OpenAlex search.
- Users evaluate randomized result layouts, select their preferred list, and optionally rank all three pipelines.
- Feedback is stored anonymously (without paper IDs) and qualitatively shows clear differences between semantic and keyword-only retrieval, guiding future ranking and query-expansion improvements.



Comparison of different seaarch methods



User preferences for different search methods in a 20-user study.

Conclusion

- This system combines multi-abstract input, semantic embeddings, and OpenAlex data API to make academic paper discovery more effective and personalized.

References

1. J. Priem et al., OpenAlex: A Fully Open Index of Scholarly Works, 2022. https://openalex.org
2. Alibaba-NLP, GTE-large-en-v1.5, Hugging Face, 2024. https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5