VECTOR DATABASES SEARCH FOR ADAPTIVE FILTERING OF SCIENTIFIC ARTICLES

Andrushchenko Dmytro,
Bachelor student of Computer Science Department
Khmelnytskyi National University

Valeriia Klimenko
Teacher of Computer Science Department
Khmelnytskyi National University

Mazurets Oleksandr
Ph.D. in Technical Sciences, Associate Professor,
Associate Professor of Computer Science Department
Khmelnytskyi National University

In the context of the rapid development of interdisciplinary research, ensuring access to high-quality and personalized scientific content is becoming especially important [1]. Traditional recommendation methods do not take into account the dynamics of user interests, the context of the query, and the complexity of scientific texts. The use of machine learning methods in adaptive filtering allows you to automate the analysis process, take into account hidden patterns in user behavior, and also increase the accuracy and relevance of recommendations [2].

The rapid and continual growth of scientific literature has made it increasingly difficult for researchers to identify, organize, and prioritize relevant publications. Traditional keyword-based search and static filtering systems struggle to cope with the volume, diversity, and semantic complexity of modern academic outputs. In this context, the integration of neural network models with vector database architectures offers a promising approach for adaptive filtering of scientific articles [3]. By leveraging learned representations of text and efficient similarity search mechanisms, such systems can provide personalized, context-aware recommendations that evolve with users' needs and the expanding corpus of publications.

Firstly, neural networks – particularly deep learning architectures such as convolutional neural networks (CNNs) and transformer-based models (e.g., BERT, RoBERTa, and their scientific-domain adaptations like SciBERT) – have demonstrated remarkable capacity to capture semantic and syntactic nuances of textual data [4, 5]. When applied to abstracts, titles, and full texts of scientific articles, these models generate high-dimensional embeddings that encapsulate not only keyword occurrences but also latent topic structures, stylistic indicators, and domain-specific jargon. Unlike traditional bag-of-words or TF-IDF representations, neural embeddings can encode subtle relationships such as paraphrasing, synonymy, and hierarchical topic similarity [6]. Consequently, articles that share conceptual content – even if they do not contain identical keywords – can be positioned closer together in the embedding space. This

foundational capability is critical for any adaptive filtering system aiming to transcend rigid, lexical matching and accommodate the dynamic semantics of scientific discourse [7].

However, generating high-quality embeddings is only one piece of the puzzle [8]. As the number of processed documents grows into the millions or tens of millions, searching for nearest neighbors in a naïvely stored embedding space becomes prohibitively expensive [9, 10]. This is where vector databases (also referred to as vector search engines or similarity search indexes) become indispensable [11]. Vector databases are specially engineered to store large sets of high-dimensional vectors and execute approximate nearest neighbor (ANN) queries with sublinear latency. Techniques such as hierarchical navigable small-world graphs (HNSW), locality-sensitive hashing (LSH), product quantization (PQ), or inverted file (IVF) structures enable real-time similarity searches even when managing billions of vectors [12].

The synergy between neural network encoders and vector databases paves the way for truly adaptive recommendation pipelines. In practice, a researcher's explicit preferences (e.g., bookmarked papers, author lists, or domain keywords) and implicit feedback (e.g., click-through history, dwell time on abstracts, or citation patterns) can be continually aggregated and encoded into a single "user embedding." This embedding then serves as a query to the vector database, retrieving candidate articles whose embeddings lie within a defined similarity radius [13]. Since both the user embedding and the article embeddings inhabit the same latent space, the retrieved set inherently reflects semantic closeness in terms of topical relevance, methodological similarity, and terminological affinity.

Another critical advantage of this approach lies in its adaptability. Neural network-based encoders can be periodically fine-tuned on new data – such as evolving research trends or emerging domain-specific jargon – ensuring that the embedding space remains aligned with the current scientific landscape. Concurrently, vector databases can be rebalanced, reindexed, or incrementally updated without interrupting the overall service, preserving low-latency query performance. This adaptability mitigates the risk of recommendation staleness, which is a common problem in conventional filtering systems where vocabularies and taxonomies can rapidly become outdated. In settings where multidisciplinary convergence (e.g., bioinformatics, computational social science) is accelerating, the ability of neural embeddings to generalize across domains is particularly valuable, allowing users to discover relevant work in adjacent fields that might not have been accessible through keyword-centric searches [14].

Despite these benefits, several challenges remain in implementing a robust adaptive filtering system. First, training or fine-tuning large neural models requires substantial computational resources – both in terms of GPUs for batch processing and skilled personnel to curate training corpora and adjust hyperparameters. Second, ensuring that embeddings accurately capture the nuances of highly specialized subfields (e.g., quantum computing or deep-sea oceanography) may necessitate domain-specific pretraining or the incorporation of specialized ontologies. Third, vector database performance can degrade if not carefully tuned; choice of distance metric (e.g., cosine similarity vs. Euclidean distance), index parameters (e.g., number of HNSW layers, PQ

code size), and update strategies (e.g., synchronous vs. asynchronous indexing) all influence retrieval accuracy and query latency. Finally, there are considerations around interpretability – researchers may wish to understand why a particular article was recommended, but the black-box nature of deep embeddings can obscure this rationale. Hybrid approaches that augment neural recommendations with transparent metadata (e.g., showing matching keywords or shared references) can help address this concern [15].

In summary, the fusion of neural network-based text encoders with vector database architectures provides a state-of-the-art framework for adaptive filtering of scientific articles. Neural models equip the system with semantically rich representations that transcend surface-level keyword matches, while vector databases ensure scalable, low-latency retrieval from massive corpora. The resulting pipelines can dynamically adapt to user preferences, incorporate continuous feedback, and remain aligned with rapidly evolving research frontiers. Although challenges related to computational cost, domain-specific nuance, and interpretability persist, ongoing advances in efficient neural architectures (such as lightweight transformers) and optimized ANN index structures promise to lower these barriers. As scientific output continues its exponential trajectory, adaptive filtering systems combining neural embeddings with vector search will be indispensable tools for researchers striving to stay informed, collaborate effectively, and discover novel insights.

Thus, improving the algorithms for personalized recommendation of scientific articles is of significant practical importance for increasing the efficiency of scientific activity and the accessibility of knowledge in the modern information environment.

The aim of the work is to research the approach to using vector databases for adaptive filtering of scientific articles.

The approach to using vector databases for adaptive filtering of scientific articles is designed for automated selection of relevant scientific publications according to the individual information needs of the user. The method helps to improve the process of adaptive filtering of scientific articles for personalized recommendations using machine learning.

The design architecture of the intelligent system for intelligent filtering of scientific articles is presented in Figure 1. The architecture of the intelligent system is based on a modular approach, where each component performs its own functional role, but they are all interconnected.

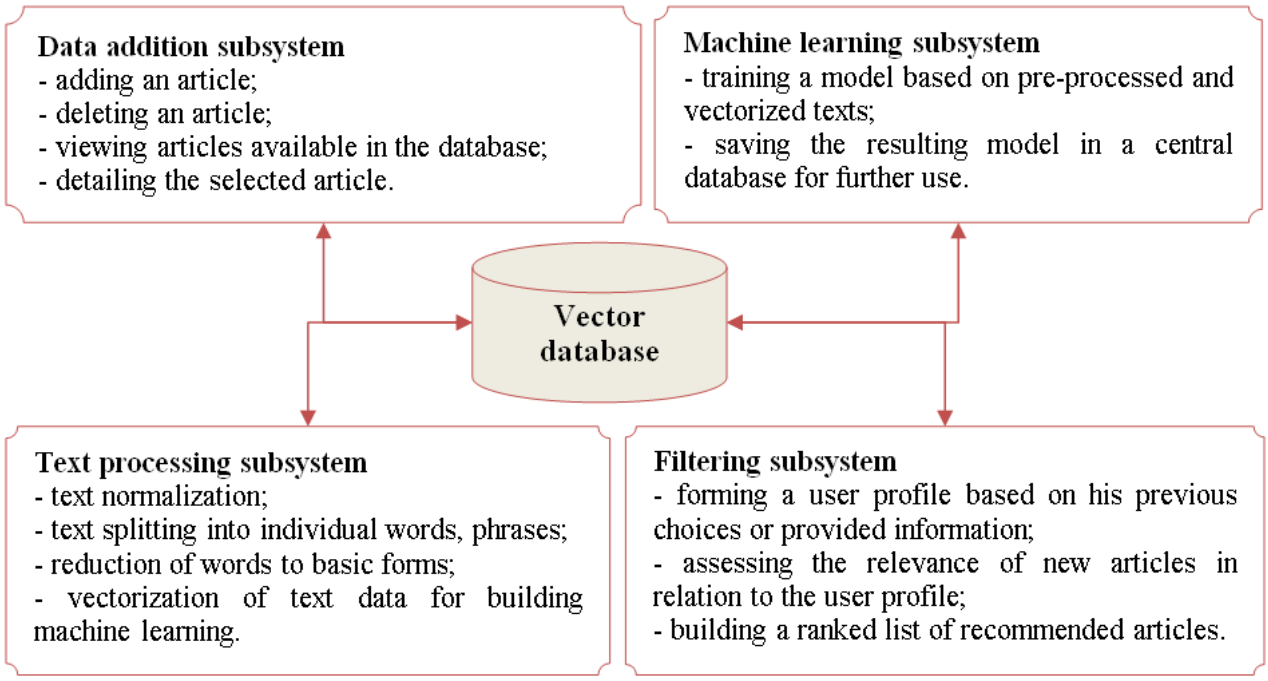

Figure 1. Architecture of intelligent system for adaptive filtering of scientific articles using vector database.

The central element of the architecture is a vector database, which serves as a repository for all text representations of scientific articles converted into multidimensional vectors. A feature of the developed system is the use of vector representation of texts instead of the traditional approach based on storing raw text. In classical search systems, each query requires reprocessing of text data, which leads to a high load on the system and limits the possibilities of in-depth content analysis. Instead, in the proposed architecture, texts are converted into vectors and mathematical objects that store the main semantic characteristics of text material in the form of coordinates in multidimensional space. Thanks to this approach, the system can compare not the texts themselves, but their vector images, which significantly speeds up the search process and allows you to assess the content similarity between documents. This means that the user receives personalized recommendations not only based on the coincidence of specific words or phrases, but also on the deep similarity of ideas, topics and context of different scientific articles.

One of the first stages of the system's functioning is the work of the data collection and management subsystem. Its main role is to organize the regular receipt of new scientific articles from various sources, such as open scientific repositories, electronic libraries and specialized scientific platforms. The functionality of the subsystem includes several important operations. First of all, it provides the ability to add new articles to the database. Adding involves automatic processing of the article text. In addition, the subsystem implements the function of deleting articles from the database. If certain records lose their relevance, contain errors or do not meet the requirements for relevance, they can be deleted to maintain the integrity and quality of the system's information environment. Another important function is viewing the articles available in the database. Users or administrators can navigate the database, see a list of available

vectorized documents, as well as their main characteristics, which allows you to control the process of filling the database.

After data collection, the materials enter the text processing subsystem. Here they are transformed for further analysis using machine learning. Text processing consists of several consecutive stages. First, the text is cleaned of extraneous characters, extra spaces, and the language of the text is normalized by reducing words to the basic form. Then the text is divided into individual elements, such as words or phrases.

At the next stage, each scientific article is processed by a pre-trained model for converting text into numerical vectors, as a result of which the text is converted into a multidimensional vector that compactly reflects its content and is suitable for further processing and analysis.

The resulting vectors are stored in a database, the structure of which is optimized to ensure fast search using the vector proximity principle. Each vector is accompanied by relevant metadata such as a unique article identifier, its title, authors' names, publication date, and a short abstract. The presence of metadata allows for effective identification of results after performing a semantic search (Figure 2).

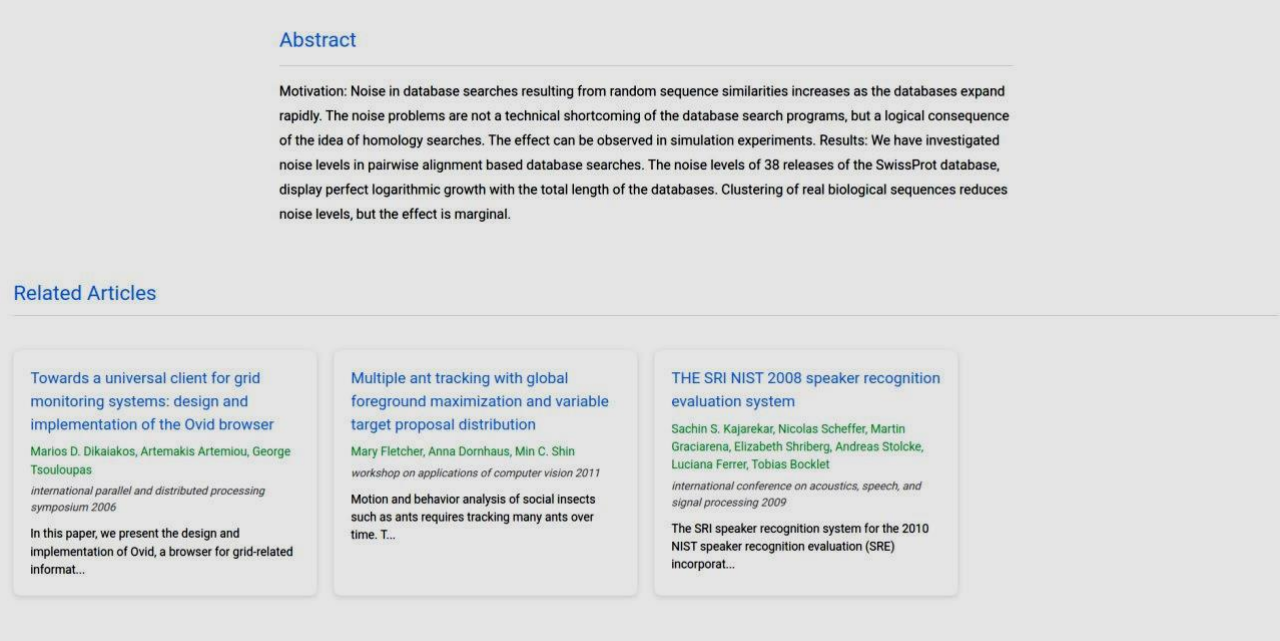

Figure 2. Results sample of adaptive filtering of scientific articles using vector DB.

Simultaneously with the processes of text collection and processing, a subsystem responsible for training a machine learning model operates. The subsystem creates a model that can effectively predict how well new articles match a user's query or interest profile. Initially, the model is trained on the basis of texts that have been pre-processed and presented in the form of vectors, as well as on examples showing which articles were relevant to the user. For this, the technology of constructing a vector user profile is used, which allows creating an average vector representation of all articles that the user found interesting. Further, when the system starts working, model continues to learn, processing new data, which allows it to adapt to changes in user preferences. The subsystem stores intermediate versions of models, compares their effectiveness on test

samples, and selects optimal options for use in a productive environment. This allows to ensure high quality personalized recommendations even in cases of sudden changes in subject of queries or user profiles.

When the user profile is formed, the filtering subsystem begins to search for relevant articles. The system searches the vector database, finding the vectors closest in semantic distance. For this, distance metrics are used, in particular cosine similarity, Euclidean distance, or more complex differentiated metrics. The search is organized using optimization structures, such as nearest neighbor search trees, which significantly speeds up the process when working with large amounts of data.

The result of the filtering subsystem is the formation of a personalized list of articles ranked by relevance. The user receives not only direct recommendations, but also the opportunity to view related information.

Therefore, vector search technologies based on neural embeddings are becoming a key tool for adaptive filtering of scientific articles. They provide flexibility, speed, and semantic accuracy of literature selection, significantly increasing the efficiency of researchers in the modern information environment. Unsolved problems, in particular in the field of interpretability and resource efficiency, open up new opportunities for research initiatives that will contribute to the further improvement of recommendation systems in the scientific field.

References:

1. Adaptive filter. Sciencedirect. https://www.sciencedirect.com/topics/computer-science/adaptive-filter
2. Using vector databases in generative artificial intelligence. Epam. https://careers.epam.ua/blog/using-vector-databases-in-generative-ai
3. Hardysh, D., Mazurets, O., & Tyschenko, O. (2024). Datalogic Relation Model for Automated Evaluating the Semantic Integrity of Test Tasks Sets by Machine Learning Means. In Innovative Solutions in Science: Balancing Theory and Practice. Proceedings 2nd International Scientific and Practical Conference (pp. 114–125).
4. Mazurets, O., & Ovcharuk, O. (2024). Efficiency Research of Method for Detecting Mental Disorders by Analysis of User Content. In Information Technology and Implementation (Satellite). Proceedings 11th International Conference (pp. 46–47).
5. Mazurets, O., & Vit, R. (2024). Practical application of method of thematic classification of text information using LDA. In Information Technology and Implementation (Satellite): Proceedings of the 11th International Conference (pp. 151–152).
6. Mazurets, O., Tymofiiev, I., & Dydo, R. (2024). Approach for Using Neural Network BERT-GPT2 Dual Transformer Architecture for Detecting Persons Depressive State. In Ricerche scientifiche e metodi della loro realizzazione: esperienza mondiale e realtà domestiche. Raccolta di articoli scientifici con gli atti della VI Conferenza scientifica e pratica internazionale (pp. 147–151).
7. Hladun, O., Mazurets, O., Molchanova, M., & Sobko, O. (2024). Real Time Detection the Person Emotion State Using Neural Network. In Scientific Research:

Modern Innovations and Future Perspectives. Proceedings of the 2 International scientific and practical conference (pp. 119–123).

8. Yurchenko, D., Mazurets, O., Didur, V., & Molchanova, M. (2024). Approach to Using Cloud Services for Visual Analytics of Neural Network Analysis of Texts Emotional Tonality. In The Future of Scientific Discoveries: New Trends and Technologies. XLVII International scientific and practical conference. (pp. 108–113).

9. Tymofiiev, I., Mazurets, O., Hardysh, D., & Molchanova, M. (2024). Neural Network Dual Architecture for Depression Detection Using Cloud Services. In Scientific Research in the Era of Digital Technologies: Challenges and Opportunities XLVI International scientific and practical conference (pp. 84–88).

10. Blazhuk, V., Mazurets, O., & Zalutska, O. (2024). An Approach to Using the mBERT Deep Learning Neural Network Model for Identifying Emotional Components and Communication Intentions. In The Impact of Scientific Research on the Development of the Modern World. Proceedings of the XLIV International Scientific and Practical Conference (pp. 79–84).

11. Mazurets, O., Sobko, O., Molchanova, M., Zalutska, O., & Yurchak, A. (2024). Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. In Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference Scientific Review of the Actual Events, Achievements and Problems (pp. 160–167). International Center of Scientific Research.

12. Mazurets, O., Sobko, O., Vit, R., & Pasternak, V. (2024). Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database. In Proceedings of XXIV International Scientific and Practical Conference Modern Scientific Challenges are the Driving Force of the Development of Scientific Research. International Scientific Unity (pp. 91–96).

13. Molchanova, M., Mazurets, O., Sobko, O., & Boiarchuk, I. (2024). Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP. In Proceedings of XXI International Scientific and Practical Conference Scientific Achievements and Innovations as a Way to Success (pp. 73–77).

14. Sobko, O., Mazurets, O., Didur, V., & Chervonchuk, I. (2024). Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior of Individuals by Text Posts. In Theoretical and Practical Aspects of Modern Research. International Scientific Unity (pp. 113–117).

15. Mazurets, O., Molchanova, M., Klimenko, V., & Prosvitliuk, M. (2024). Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. In Prospects of Scientific Research in the Conditions of the Modern World (pp. 97–102).