Multifactor vectorized search for relevant scientific articles: an adaptive deep learning pipeline method

Olexander Barmak[1,†], Olexander Mazurets[1,†], Olena Sobko[1,*,†], Dmytro Andrushchenko[1,†] and Iurii Krak[2,3,†]

[1] *Khmelnytskyi National University, 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, Kyiv, 64/13, Volodymyrska str., 01601, Ukraine*
[3] *Glushkov Cybernetics Institute, Kyiv, 40, Glushkov ave., 03187, Ukraine*

Abstract

In paper proposed approach to solving the problem of filtering scientific information in educational process, taking into account not only the content aspect of publications, but also the integration of structural meta-attributes, such as chronology, citations, publication context, etc. As part of the study, a multifactorial method for vectorized search of scientific publications was designed and evaluated, combining the MiniLM-L6-v2 transformer model for constructing a latent semantic representation with adaptive ranking of results by meta-attributes. A distinctive feature of the approach is the integration of content and structural relevance, which provides increased accuracy of retrieval in a large-scale academic environment. An experimental study conducted on an open dataset with over a million scientific articles confirmed the superiority of the developed method over basic approaches. The values of the metrics Semantic Precision@10 = 0.70, nDCG@10 = 0.73, MRR@10 = 0.61 indicate the relevance, stability and efficiency of search results. Compared to modern analogues, the method demonstrated an improvement of 5-15\% depending on the metric. The obtained results indicate the feasibility of using compact models such as MiniLM in combination with multi-criteria ranking in academic search tasks. The proposed method showed high scalability and potential for implementation in educational information systems and academic digital library systems.

Keywords

multifactor vectorized search, semantic search, deep learning, MiniLM-L6-v2

1. Introduction

In today's rapidly growing volume of scientific publications, there is an urgent need for effective tools for semantic search and information filtering in the educational process [1, 2]. Traditional approaches to information search, based mainly on keywords [3], are insufficient to accurately establish the content relevance between a query and a publication [4]. This is especially noticeable in fields with high interdisciplinarity, where formal terminology often varies [5]. In response to these challenges, new generation search engines are gradually integrating deep learning methods that provide the construction of vector representations of texts capable of capturing latent semantic dependencies [6]. The relevance of the study is due to the need to create adaptive systems for filtering scientific information that not only take into account the content aspect of publications, but also integrate structural meta-attributes, such as chronology, citations, and context of the publication.

The main contributions of the study are the combination of vector search based on embeddings with meta-ranking of results by structural features. This combination allows to ensure not only the content correspondence to the search query, but also to take into account the qualitative and

IntelITSIS-2026: 7th International Workshop on Intelligent Information Technologies & Systems of Information Security, July 3, 2026, Khmelnytskyi, Ukraine
[1,*] Corresponding author.
[†] These authors contributed equally.
alexander.barmak@gmail.com (O. Barmak); exe.chong@gmail.com (O. Mazurets); olenasobko.ua@gmail.com (O. Sobko); andrushchenkodmytro560@gmail.com (D. Andrushchenko); yuri.krak@gmail.com (I. Krak)
0000-0003-0739-9678 (O. Barmak); 0000-0002-8900-0650 (O. Mazurets); 0000-0001-5371-5788 (O. Sobko); 0009-0003-8861-5922 (D. Andrushchenko); 0000-0002-8043-0785 (I. Krak)

temporal characteristics of publications, which is especially relevant in the educational process for the analysis of modern scientific information.

2. Related works

The aim of this literature review is to explore existing approaches to vectorized semantic search, identify their advantages and limitations.

Paper [7] proposes a solution is described that organizes items into semantically consistent clusters and uses a search engine to increase the diversity of recommendations. The system transforms items into vector form and clusters them according to semantic similarity through an online algorithm with an adaptive threshold. The mechanism, which can be controlled by the user, selectively shows items from less studied groups, due to which the recommendations become more diverse. Experiments on the MovieLens dataset [8] show that the inclusion of search reduces the similarity of items in the list from 0.34 to 0.26 and increases the surprise to 0.73. This gives better results than traditional methods of collaborative filtering and popularity recommendations.

The researchers in [9] focused on creating an intelligent recommendation system based on self-attention mechanisms. They emphasize the importance of deep analysis of user behavior, including interaction with previous recommendations. The proposed model demonstrated high accuracy rates of up to 85.2% and -measure of 84.3%, which indicates its prospects for implementation in scientific systems.

It is worth noting the study [10], which considers an adaptive algorithm for collaborative filtering taking into account the fading of interest over time. This approach makes it possible to effectively solve the problem of «cold start» and adapt recommendations in accordance with changes in the user's scientific activity.

Also, in [11], a hybrid system combining the advantages of content and collaborative approaches is considered. The main attention is focused on developing a model that takes into account multidimensional characteristics of articles, in particular metadata, semantic embeddings and viewing history. And in [12] a system for interactive search of scientific publications is presented, which combines semantic analysis, Knowledge Graphs and LLM. The system implements semantic search to obtain a relevant corpus of scientific articles, from which meaningful information is automatically extracted. The system provides the user with structured answers based on identified scientific statements.

A ranking system for improving the search of biomedical datasets is presented in [13], which combines query expansion using word embeddings, a similarity measurement algorithm taking into account the relevance of terms, and categorization of results. The system was tested on a corpus of 800 thousand datasets and 21 queries, showing competitive results – in particular, the highest infAP value of +22.3%. The query expansion module made the greatest contribution to increasing accuracy, while categorization did not have a significant effect. The developed solution is considered a promising alternative to traditional terminological approaches, although it requires further research.

The authors of [14] presented the NLP-KG system, designed to facilitate the search and study of scientific literature in the field of natural language processing, especially in situations where the user has no previous experience in the subject. The system combines semantic search, review article filtering, a Fields of Study graph, and an interactive LLM-based interface that allows you to ask questions about unfamiliar concepts or individual publications. NLP-KG creates knowledge based on scientific articles from the ACL Anthology and arXiv cs.CL corpus, demonstrating high results in classification and search metrics. The main limitations of the system are the subjectivity of expert composition of the field hierarchy graph and limited coverage of the literature, which does not include conferences such as AAAI, NeurIPS, or ICLR.

Knowledge Navigator is proposed in [15]– a scientific literature research system that combines semantic search methods with the possibility of thematic review of results. The system is based on

a two-level hierarchy of topics and subtopics, which is automatically built based on user query processing.

Unlike the reviewed systems, the proposed approach combines compact transformer-based semantic embeddings with explicit metadata-aware re-ranking within a scalable FAISS retrieval pipeline. In contrast to existing solutions focused primarily on semantic retrieval or recommendation, the proposed method integrates publication quality indicators while preserving computational efficiency on large-scale academic collections, making it suitable for educational information retrieval.

Thus, modern research in the field of adaptive filtering of scientific articles confirms the effectiveness of deep learning methods for searching and filtering scientific publications.

Accordingly, the research purpose is to develop and substantiate an adaptive pipeline method using deep learning, which combines vectorized representation of queries and texts and multifactor ranking of results taking into account structural meta-attributes of publications.

3. Method design

The proposed method (Figure 1) is based on a multi-stage approach to semantic retrieval of scientific articles, which integrates vector representations of text objects, multifactor ranking of results and adaptive weighting of criterion features. The goal is to formalize a relevance function that combines latent semantic similarity between a query and a document with additional meta-features that characterize the quality and relevance of a scientific work.

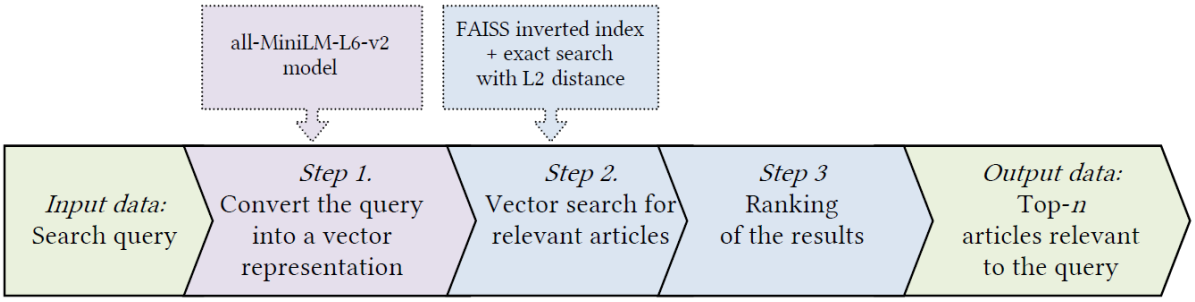

Figure 1: Scheme of adaptive pipeline method using deep learning for multifactor vectorized search for relevant scientific articles.

Let the user's query q be an aggregated textual construct that includes key information components: title, annotation, keywords, authors and, if necessary, other relevant metadata. Similarly, each document $d_i \in D$ is described by a set of fields:

$$d_i = \{t_i, a_i, k_i, u_i, v_i^{(meta)}\}, \quad (1)$$

where t_i – the publication title, a_i – the abstract, k_i – the keywords, u_i – the authors, and $v_i^{(meta)}$ – other text meta fields (if any).

Thus, the generalized query looks like this:

$$q = q(t) \oplus q(a) \oplus q(k) \oplus q(u), \quad (2)$$

where $q(t)$ is the search query component containing the publication title, $q(a)$ is the search query component containing the annotation, $q(k)$ is the search query component containing the keywords; $q(u)$ is the search query component containing the authors.

The aggregated query is passed to the vector representation model F, which is implemented based on a transformer architecture, in particular a Sentence-BERT type model. As a result, the query vector v_q is calculated, which reflects its latent semantics:

$$v_q = F(q). \tag{3}$$

Similarly, representation 2 provides vectorization of all documents with which comparison will be made.

The implementation of function F uses the pre-trained all-MiniLM-L6-v2 model, which provides a compact and efficient vector representation for short texts and metadata. The choice of such architecture is due to special optimization for semantic search tasks and creation of «sentence embeddings». The architecture of the used vectorization model is shown in Figure 2.

The initial stage of text processing in the proposed system involves converting the input sentence into a sequence of tokens using a tokenizer of the WordPiece type, which is an integral component of models built on the BERT architecture [16]. This tokenizer implements sublinguistic text segmentation, allowing for efficient work with new or morphologically complex words, ensuring the constancy of the dictionary dimension with a high coverage of lexemes [17].

During the tokenization process, special service tokens are added to the beginning and end of the sequence: [CLS], which serves as a global representation for the entire sequence, and [SEP], which performs the function of a sentence separator or a marker for the end of the input text. After tokenization, each element is passed to the embedding layer, where three components are combined: token embeddings, positional embeddings, and segmental embeddings. Token embeddings determine the semantic meaning of a particular word, positional embeddings add information about the order of tokens, and segmental embeddings are used to differentiate between parts of the text in tasks such as next sentence prediction.

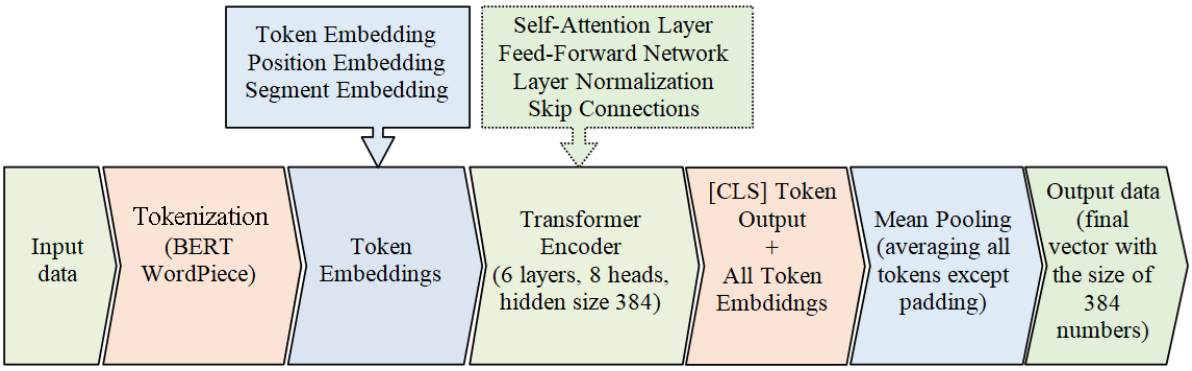

Figure 2: Neural network architecture.

The implementation of the all-MiniLM-L6-v2 model [18], which belongs to the Sentence-BERT model class [19], provides six serially connected transformer blocks. Each block consists of two main components: a Multi-Head Self-Attention layer and a two-layer fully connected Feed-Forward Network, supplemented by the Residual Connection and Layer Normalization mechanisms. The multi-head self-attention mechanism is implemented through eight parallel projection spaces, which allows the model to simultaneously focus on different aspects of the context and detect both local and long-term dependencies between tokens.

After each layer of attention, the result is passed to a feed-forward network with nonlinear activation of the GELU or ReLU type, which allows the model to adaptively transform vector representations. Such a sequence of transformer blocks provides progressive refinement of the semantic representation of each token, taking into account the context of the entire sequence.

At the output, each token is represented by a contextualized vector of fixed dimension, however, to form a single semantic representation of the entire sequence (sentence or query), the

mean pooling strategy is used, which calculates the average value of the vectors of all tokens, excluding service ones. As a result, a sentence-level embedding of dimension 384 is formed, which is fed to the input of subsequent modules for calculating semantic similarity.

To perform an effective search in the space of vector representations, the FAISS system [20] is used, which provides scalable indexing based on inverted structures and accurate search by the Euclidean metric L_2-norm. That is, first a subset of potentially relevant documents is formed through rapid approximate selection, after which they are further ranked according to the integral relevance function:

$$R(d_i) = \alpha sim(q, d_i) + \beta y_i + \gamma c_i + \delta v_i, \quad (4)$$

where $sim(q, d_i)$ – cosine similarity value of vectorized query q to vectorized document d_i; $y_i \in [0,1]$ – normalized value of publication year; $c_i \in [0,1]$ – normalized number of citations; $v_i \in [0,1]$ – numerical representation of prestige of place of publication, which is based on quartiles; $\alpha + \beta + \gamma + \delta = 1$ – weight coefficients, which specify the influence of each component on the final ranking. In the experiments, the weighting coefficients were empirically selected as α = 0.70, β = 0.10, γ = 0.10 and δ = 0.10, giving priority to semantic similarity while preserving the influence of metadata. The publication year and citation count were normalised using min–max normalisation over the dataset. Venue prestige was encoded according to journal quartiles (Q1 = 1.0, Q2 = 0.75, Q3 = 0.50, Q4 = 0.25, other venues = 0). FAISS indexing was implemented using the IndexFlatIP configuration with cosine similarity computed on L2-normalised embeddings.

The proposed multi-stage method of vectorized search of scientific articles implements a combination of deep learning transformer models with indexing and multifactor ranking mechanisms. Due to the aggregated representation of the query and documents, which takes into account both textual and structural information, increased sensitivity to semantic correspondence is provided. Vectorization based on Sentence-BERT allows you to form compact latent representations, convenient for scalable search in vector space using FAISS tools. The introduced relevance function integrates semantic similarity with meta-attributes of publication quality, which makes it possible to adapt the system to the specifics of the subject area or the needs of end user.

4. Experimental research

4.1. Dataset

To test the developed adaptive pipeline method using deep learning for multifactor vectorized search for relevant scientific articles, the Research Papers Dataset [21] dataset hosted on the Kaggle platform [22] was used. The proposed dataset covers articles from various fields of research, including computer science, mathematics, physics, etc.

The dataset consists of records representing scientific articles with the following key fields: unique identifier "id", article title "title", abstract «abstract», authors "authors", publication location "venue", "year", number of citations "n_citation" and list of references "references". The "title" field is used as an indicator of relevance and a source of key terms for vector representation, and "abstract" is used as the main text field for deep semantic analysis. Additional metadata ("authors", "venue", "year", "n_citation") provide opportunities for filtering, chronological analysis, assessment of publication authority and construction of citation networks. To achieve symmetry in the representation, each class contains the same number of elements, which ensures a reduction in model bias during training. In total, the dataset contains about a million scientific articles, making it suitable for modeling scenarios close to the real conditions of scaled academic search.

4.2. Experimental software

The intelligent system based on the developed method is implemented as a web application with client and server parts. The main page performs the main function of user interaction with the system. It contains a search form and a selector for choosing which criteria to search by. The form also allows entering keywords or using voice search. The developed intelligent system is shown in Figure 3.

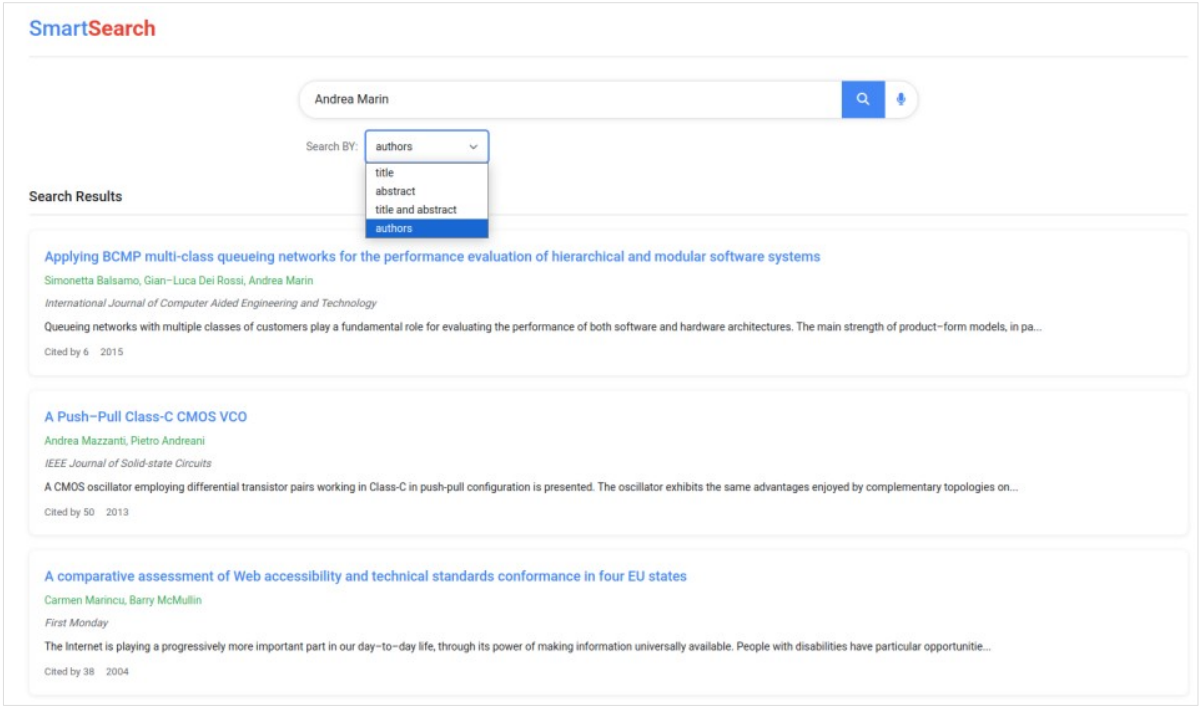

Figure 3: Screenshot of the developed web application.

The interface allows users to specify search criteria, submit text or voice queries and inspect ranked search results together with semantic relevance and metadata. The ranking output combines semantic similarity with publication year, citation count and venue prestige, enabling transparent interpretation of retrieval results.

The functioning of the developed intelligent system uses a microservice architecture. Each service is responsible for a certain logic, which allows for system scaling [23]. The architecture of the web application is schematically presented in Figure 4.

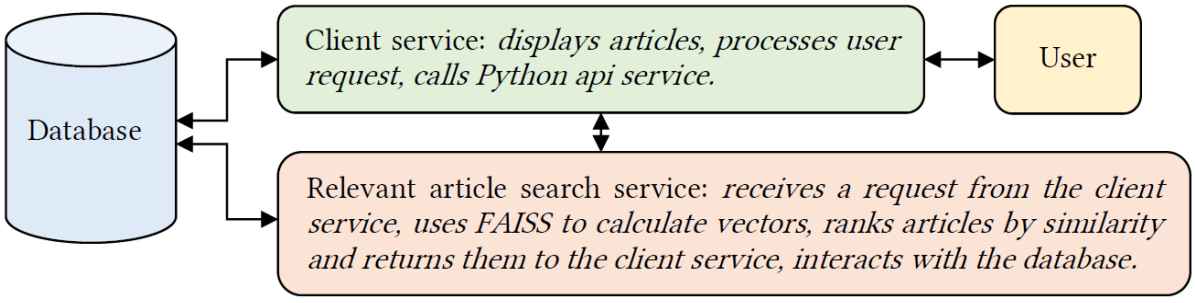

Figure 4: Microservices architecture of the web application.

In the proposed web application architecture, the entry point is a client service that provides a graphical user interface, processing of search queries, and interaction with the server part via RESTful API [24]. The user can initiate a search for scientific publications using text or voice input, after which the query is passed to the relevant document search module.

The server search module implements the proposed adaptive pipeline method, which transforms the aggregated search query into a vector representation v_q using the Sentence-BERT transformer

model. All scientific articles in the database are pre-vectorized in a similar latent space, which allows for semantic matching between the query and documents.

FAISS indexing [25] is used to search for vectors, which implements a scalable inverted structure with support for accurate calculation of the cosine distance. Based on the calculated semantic similarity and meta-parameter values, multifactor ranking of results is performed according to the relevance function 4. The generated list of relevant documents is passed back to the client service in a structured JSON format for further display in the user interface.

5. Result and discussion

To evaluate the effectiveness of the proposed method, a series of experiments were conducted on the open dataset Research Papers Dataset, which contains text and meta-information fields of scientific publications from various fields. The comparison was made between traditional approaches [26], BM25 [27]), neural network models (Universal Sentence Encoder [28] and the developed multi-factor method based on MiniLM-L6-v2.

The experimental methodology involved the formation of 100 random aggregated queries, which were manually marked by three IT experts. The experts determined the relevance of potentially related articles based on the topic, content, keywords and industry affiliation. The level of inter-expert agreement was k = 0.82 according to the Fliess scale [29], which indicates a high agreement of the estimates. The queries were generated by randomly selecting publications from different scientific domains and constructing aggregated search requests using their titles, abstracts and keywords. Three independent experts in information technologies assessed relevance on a binary scale. In cases of disagreement, the final label was assigned by majority voting. The experts were not informed which retrieval method produced a given ranked list, and all compared methods were evaluated using the same candidate document pool.

Based on this markup, the system returned the top 10 results for each search method for each query. The metrics Semantic Precision@10, nDCG@10, and MRR@10 were calculated automatically using open libraries [30] by comparing the ranked lists with previously marked-up relevant documents. The values of the obtained metrics are given in Table 1.

Table 1
Comparison of the known methods with the developed one

Search method	Semantic Precision@10	nDCG@10	MRR@10
TF-IDF method	0.41	0.43	0.32
BM25 method	0.46	0.48	0.37
Universal Sentence Encoder method	0.59	0.62	0.50
MiniLM-L6-v2 method (without metadata)	0.65	0.68	0.56
Proposed method (MiniLM-L6-v2 + metadata)	0.70	0.73	0.61

The results obtained demonstrated the superiority of the proposed method: the highest values of the metrics were achieved due to the combination of a compact and efficient vector representation (MiniLM-L6-v2) with the use of structural metadata (year, citations, prestige of the publication). Compared to the base model without metadata, the indicators improved by an average of 5–6%, and compared to TF-IDF – by about 50% for MRR@10.

The value of nDCG@10 = 0.73 indicates that relevant documents are not only present among the results, but also occupy higher positions in the output, which is critically important for academic search systems. The model also demonstrated good performance and scalability, which makes it suitable for practical implementation in digital libraries and scientific information platforms.

The result was also compared with a similar study conducted by the authors [13], where the nDCG value of 0.5905 was achieved. Compared to the obtained values, the developed approach demonstrates higher results (0.73).

Among the potential limitations, it should be noted the dependence of the ranking quality on the completeness and accuracy of metainformation. In particular, citations and publication quartile. In addition, the preliminary vectorization of a large corpus of documents requires additional computational resources at the indexing stage. Also, the MiniLM-L6-v2 model, although optimal for performance, may be inferior to deeper transformer models in the nuances of semantics.

In further research, it is planned to expand the method by integrating the citation context, adapting to the educational process, and involving personalized user profiles. This will allow not only to analyze the relevance of individual articles, but also to build dynamic trajectories of scientific search depending on the interests of the researcher.

6. Conclusion

This paper presented an adaptive deep learning pipeline for multifactor vectorized retrieval of scientific articles that combines semantic text representations with metadata-aware ranking. The proposed approach integrates MiniLM-L6-v2 embeddings, scalable FAISS indexing, and a relevance function that jointly considers semantic similarity, publication year, citation count, and publication venue prestige.

The experimental evaluation demonstrated that incorporating structural metadata into semantic retrieval improves the quality of ranked search results compared with both traditional lexical retrieval methods and semantic retrieval without metadata-aware re-ranking. The obtained results confirm the effectiveness of combining compact transformer-based embeddings with multifactor ranking for large-scale academic search tasks.

The proposed method is computationally efficient, scalable, and suitable for integration into educational information systems and academic digital libraries. Future work will focus on extending the ranking model by incorporating citation context, user personalisation, and dynamic researcher profiles to further improve retrieval quality and support adaptive scientific information discovery.

Acknowledgements

The authors express their gratitude to Prof. Tetiana Hovorushchenko, Prof. Oleh Savenko, Prof. Sergii Lysenko and Dr. Peter Popov to scientific and organizational support of 7th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2026).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] N. Muhammadjon, Integration of electronic thesauri with search systems, information retrieval and natural language processing, in: International Conference on Modern Development of Pedagogy and Linguistics, volume 2, 2025, pp. 24–26. URL: https://universalconference.us/universalconference/index.php/icmdpl/article/view/3778.

[2] R. Kumar, S. Sharma, Hybrid optimization and ontology-based semantic model for efficient textbased information retrieval, The Journal of Supercomputing 79 (2023) 2251–2280. doi:10.1007/s11227-022-04708-9.
[3] K. Lipianina-Honcharenko, V. Vitenko, D. Zahorodnia, Tools for semantic search and answer generation in Ukrainian-language museum systems, Computer Systems and Information Technologies (CSIT) (2026) 70–77. doi: 10.31891/csit-2026-2-7.
[4] Z. Zhang, B. G. Patra, A. Yaseen, J. Zhu, R. Sabharwal, K. Roberts, T. Cao, H. Wu, Scholarly recommendation systems: a literature survey, Knowledge and Information Systems 65 (2023) 4433–4478. doi:10.1007/s10115-023-01901-x.
[5] O. V. Barmak, O. V. Mazurets, I. V. Krak, A. I. Kulias, L. E. Azarova, K. Gromaszek, S. Smailova, Information technology for creation of semantic structure of educational materials, in: Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, volume 11176, 2019, pp. 616–626. doi:10.1117/12.2537064.
[6] I. Pinedo, M. Larrañaga, A. Arruarte, Arzigo: A recommendation system for scientific articles, Information Systems 12 (2024) 102367. doi:10.1016/j.is.2024.102367.
[7] E. Bianchi, Beyond relevance: An adaptive exploration-based framework for personalized recommendations, arXiv preprint arXiv:2503.19525 (2025). URL: https://arxiv.org/abs/2503.19525.
[8] G. Aridor, D. Gonçalves, R. Kong, D. Kluver, J. A. Konstan, The MovieLens Beliefs Dataset: Collecting Pre-Choice Data for Recommender Systems, Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 1. doi:10.1145/3640457.368815.
[9] H. Zhou, Intelligent personalized content recommendations based on neural networks, International Journal of Intelligent Networks 4 (2023) 231–239. doi:10.1016/j.ijin.2023.09.001.
[10] A. Ghiye, B. Barreau, L. Carlier, M. Vazirgiannis, Adaptive collaborative filtering with personalized time decay functions for financial product recommendation, in: Proceedings of the 17th ACM conference on recommender systems, 2023, pp. 798–804. URL: https://arxiv.org/abs/2308.01208.
[11] Y. Afoudi, M. Lazaar, M. A. Achhab, Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network, Simulation Modelling Practice and Theory 113 (2021) 102375. doi:10.1016/j.simpat.2021.102375.
[12] A. Oelen, M. Y. Jaradeh, S. Auer, Orkg ask: a neuro-symbolic scholarly search and exploration system, arXiv preprint arXiv:2412.04977 (2024). URL: https://arxiv.org/abs/2412.04977.
[13] D. Teodoro, L. Mottin, J. Gobeill, A. Gaudinat, T. Vachon, P. Ruch, Improving average ranking precision in user searches for biomedical research datasets, Database (2017). doi:10.1093/database/bax083.
[14] T. Schopf, F. Matthes, Nlp-kg: A system for exploratory search of scientific literature in natural language processing, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2024, pp. 127–135. doi:10.18653/v1/2024.acl-demos.13.
[15] U. Katz, M. Levy, Y. Goldberg, Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 8838–8855. doi:10.18653/v1/2024.findings-emnlp.516.
[16] Z. Feng, D. Tang, X. Feng, C. Zhou, J. Liao, S. Wu, B. Qin, Y. Cao, S. Shi, Pretraining without 8 wordpieces: learning over a vocabulary of millions of words, International Journal of Machine Learning and Cybernetics 15 (2024) 3989–3998. doi:10.1007/s13042-024-02132-4.
[17] I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk, O. Barmak, Method for neural network detecting propaganda techniques by markers with visual analytic, in: CEUR Workshop Proceedings, volume 3790, 2024, pp. 158–170. URL: https://ceur-ws.org/Vol-3790/paper14.pdf.

[18] sentence-transformers/all-minilm-l6-v2, 2025. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
[19] L. Stankevičius, M. Lukoševičius, Extracting sentence embeddings from pretrained transformer models, Applied Sciences 14(19) (2024) 8887. doi:10.3390/app14198887.
[20] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, arXiv preprint arXiv:2401.08281 (2024). URL: https://arxiv.org/abs/2401.08281.
[21] Research papers dataset, 2025. URL: https://www.kaggle.com/datasets/nechbamohammed/research-papers-dataset.
[22] I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Zalutska, E. Manziuk, O. Barmak, Method for political propaganda detection in internet content using recurrent neural network models ensemble, in: CEUR Workshop Proceedings, volume 3806, 2024, pp. 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf.
[23] O. öderman, Comparison of web application architecture styles, 2025. URL: https://aaltodoc.aalto.fi/server/api/core/bitstreams/4e1121c4-6409-4032-91aa-381b7b17096b/content.
[24] J. Bogner, S. Kotstein, T. Pfaff, Do RESTful API design rules have an impact on the understandability of Web APIs?, Empirical Software Engineering 28 (2023) 132. doi: 10.1007/s10664-023-10367-y.
[25] Github - facebookresearch/faiss: A library for efficient similarity search and clustering of dense vectors, 2025. URL: https://github.com/facebookresearch/faiss.
[26] Y. Krak, O. Barmak, O. Mazurets, The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials, in: CEUR Workshop Proceedings, volume 1631, 2016, pp. 237–245. URL: https://ceur-ws.org/Vol-1631/237-245.pdf.
[27] J. Lin, X. Ma, S. Lin, J. Wang, Pretrained Transformers for Text Ranking: A Survey, Foundations and Trends in Information Retrieval 16(3–4) (2022) 1–119. doi: 10.1561/1500000072.
[28] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 87–94. doi: 10.18653/v1/2020.acl-demos.12.
[29] K. Han, L. Ryu, Statistical methods for the analysis of inter-reader agreement among three or more readers, Korean Journal of Radiology 25 (2024) 325. doi:10.3348/kjr.2023.0965.
[30] A. Moffat, Batch evaluation metrics in information retrieval: Measures, scales, and meaning, IEEE Access 10 (2022) 105564–105577. doi: 10.1109/ACCESS.2022.3211668.