DOI: 10.34185/1991-7848.itmm.2026.01.094

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЛЕКСИЧНОГО ТА СЕМАНТИЧНОГО ПОШУКУ У БАГАТОМОВНИХ ВЕБСЕРВІСАХ АГРЕГАЦІЇ МЕДІАКОНТЕНТУ

Сімакін С.К. [1] [ORCID], Божуха Л.М. [2] [ORCID]

[1] Дніпровський національний університет імені Олеся Гончара, аспірант, Україна

[2] Дніпровський національний університет імені Олеся Гончара,
канд. фіз.-мат. наук, доцент, Україна

Анотація. *У роботі представлено результати порівняльного аналізу двох підходів до організації пошуку у вебсервісах агрегації багатомовного медіаконтенту: лексичного пошуку на основі алгоритму BM25 (Elasticsearch) та семантичного пошуку на основі щільних векторних вкладень моделі BGE-M3 (Qdrant). Розроблено платформу MediaAggregator, яка індексує 70 000 новинних статей сімома мовами та надає уніфікований програмний інтерфейс (REST API) для порівняння якості пошуку, латентності та обсягу відповіді. Експерименти проведено локально на процесорі AMD Ryzen 9 8945HS із 32 ГБ оперативної пам'яті з використанням ONNX Runtime для інференсу моделі вкладень на CPU. Результати демонструють, що лексичний пошук забезпечує в середньому у 2,7 рази нижчу латентність, тоді як семантичний пошук забезпечує крос-лінгвістичне знаходження релевантного контенту незалежно від мови запиту, що є критичним для оптимізації мультимовних вебсервісів.*

Ключові слова: *вебсервіс, пошукова оптимізація, семантичний пошук, Elasticsearch, векторна база даних, BGE-M3, багатомовний пошук, BM25, Qdrant.*

Вступ

Сучасні вебсервіси агрегації медіаконтенту оперують великими обсягами багатомовних даних та потребують ефективних механізмів пошуку інформації. Традиційний лексичний пошук, заснований на алгоритмі BM25, залишається стандартом індустрії завдяки високій швидкодії та передбачуваності результатів [1]. Проте він обмежений точним збігом ключових слів та не здатний враховувати семантичну близькість понять різними мовами.

З розвитком технологій великих мовних моделей (LLM) з'явилися нові підходи до пошуку, засновані на щільних векторних вкладеннях (dense embeddings). Моделі, такі як BGE-M3 від BAAI [2], підтримують понад 100 мов та

дозволяють здійснювати крос-лінгвістичний семантичний пошук, де запит однією мовою знаходить релевантні документи іншими мовами [3].

Однак практичне застосування семантичного пошуку у продуктивних вебсервісах потребує дослідження компромісів між латентністю, релевантністю та масштабованістю. Метою даної роботи є порівняльний аналіз лексичного та семантичного підходів до пошуку на реальному багатомовному корпусі новинних статей з використанням спеціально розробленої платформи оцінювання.

Основний матеріал

Архітектура платформи. Для проведення дослідження розроблено платформу MediaAggregator, яка складається з: бекенду на ASP.NET Core 10, який надає REST API з двома пошуковими ендпоінтами; Elasticsearch 9.3 для лексичного пошуку з алгоритмом ранжування BM25; Qdrant 1.17 як векторної бази даних для семантичного пошуку з косинусною подібністю; ONNX Runtime для інференсу моделі вкладень BGE-M3 (1024 виміри); React-інтерфейсу для візуального порівняння результатів.

Тестове середовище. Експерименти проведено локально на ноутбуці з процесором AMD Ryzen 9 8945HS (архітектура Zen 4, 8 ядер / 16 потоків, базова частота 4,0 ГГц, максимальна – 5,2 ГГц, 24 МБ кешу L2+L3, TDP 45 Вт, техпроцес TSMC 4 нм) та 32 ГБ оперативної пам'яті LPDDR5x. Усі компоненти системи (Elasticsearch, Qdrant, API, модель BGE-M3) розгорнуто через Docker Compose на одній машині. Інференс моделі вкладень виконувався на CPU через ONNX Runtime без GPU-прискорення.

Набір даних. Використано корпус Babel Briefings – 70 000 новинних статей (по 10 000 на мову) сімома мовами: українською, англійською, німецькою, французькою, іспанською, польською та італійською. Кожна стаття містить оригінальний заголовок та опис, а також їх англомовні переклади, що дозволяє лексичному пошуку працювати крос-лінгвістично через англомовні поля.

Методика експерименту. Виконано серію пошукових запитів різними мовами через обидва пошукові движки. Для кожного запиту зафіксовано:

латентність відповіді (мс), кількість знайдених документів, обсяг відповіді (байти), мовний склад топ-10 результатів. Результати наведено у табл. 1.

Таблиця 1

Порівняння результатів лексичного та семантичного пошуку

Запит	Мова	L_{lat}, мс	S_{lat}, мс	N_{lex}	N_{sem}	Крос.
climate change global warming	en	99,1	279,3	3 040	70 000	Так
зміна клімату глобальне потепління	uk	48,9	95,6	22	70 000	Так
artificial intelligence machine learning	en	49,2	127,5	852	70 000	Так
штучний інтелект технології	uk	24,4	104,5	453	70 000	Так

У табл. 1: Llat – латентність лексичного пошуку; Slat – латентність семантичного пошуку (включаючи генерацію вкладення); Nlex, Nsem – кількість знайдених документів; Крос. – наявність крос-лінгвістичних результатів у семантичному пошуку.

Аналіз латентності. Середня латентність лексичного пошуку склала 55,4 мс, семантичного – 151,7 мс. Семантичний пошук повільніший у середньому в 2,7 рази, що пояснюється додатковим етапом генерації векторного вкладення запиту через модель BGE-M3 за допомогою ONNX Runtime на CPU. Для англомовних запитів різниця сягає 2,8 рази (99 мс проти 279 мс), тоді як для коротших українських запитів – 2,6 рази (37 мс проти 100 мс).

Аналіз крос-лінгвістичних можливостей. Ключовою перевагою семантичного пошуку є здатність до крос-лінгвістичного пошуку. Запит українською мовою «зміна клімату глобальне потепління» через лексичний пошук знайшов лише 22 українських документи. Семантичний пошук для того ж запиту повернув релевантні результати іспанською, італійською, німецькою, польською та французькою мовами, оскільки модель BGE-M3 відображає семантично близький контент у спільний векторний простір незалежно від мови [2].

Аналіз повноти пошуку. Лексичний пошук повертає лише документи з точним збігом термінів (від 22 до 3040 результатів залежно від запиту). Семантичний пошук ранжує весь корпус (70 000 документів) за косинусною подібністю, забезпечуючи повне покриття колекції. Це має практичне значення для вебсервісів, де важливо не пропустити релевантний контент.

Оптимізація продуктивності. Для зменшення латентності семантичного пошуку реалізовано конвеєрну архітектуру: паралельна токенізація наступного запиту під час інференсу поточного, підтримка GPU-прискорення через DirectML та CUDA, пакетна обробка з сортуванням за довжиною тексту для мінімізації витрат на доповнення (padding) у ONNX Runtime [4].

Висновки

Проведений порівняльний аналіз лексичного (Elasticsearch BM25) та семантичного (Qdrant + BGE-M3) підходів до пошуку на корпусі з 70 000 багатомовних новинних статей дозволяє сформулювати такі висновки:

1. Лексичний пошук забезпечує в 2,7 рази нижчу латентність (55 мс проти 152 мс), що робить його оптимальним для сценаріїв із жорсткими вимогами до швидкодії.

2. Семантичний пошук забезпечує крос-лінгвістичне знаходження релевантного контенту, що є критичним для багатомовних вебсервісів, де користувач може формулювати запити будь-якою мовою.

3. Гібридний підхід, що поєднує обидва методи, є перспективним напрямком оптимізації вебсервісів пошуку: лексичний пошук для швидкої фільтрації, семантичний – для уточнення та розширення результатів.

Подальші дослідження спрямовані на реалізацію гібридного пошуку з адаптивним зважуванням результатів обох методів та оптимізацію інференсу вкладень на GPU.

ЛІТЕРАТУРА / REFERENCE

1. Mitra B., Craswell N. An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval. 2018. Vol. 13, No. 1. P. 1–126. DOI: 10.1561/1500000061.
2. Chen J., Xiao S., Zhang P. et al. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:2402.03216. 2024. DOI: 10.48550/arXiv.2402.03216.

3. Karpukhin V., Oguz B., Min S. et al. Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 6769–6781.
4. ONNX Runtime: cross-platform, high performance ML inferencing and training accelerator. URL: https://onnxruntime.ai/ (дата звернення: 05.04.2026).
5. Elasticsearch: The Official Distributed Search & Analytics Engine. URL: https://www.elastic.co/elasticsearch (дата звернення: 05.04.2026).
6. Qdrant – Vector Search Engine. URL: https://qdrant.tech/ (дата звернення: 05.04.2026).
7. Kamphuis C., de Vries A. P., Boytsov L., Lin J. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. Proceedings of the 42nd European Conference on Information Retrieval (ECIR). 2020. P. 28–34. DOI: 10.1007/978-3-030-45442-5_4.
8. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP). 2019. P. 3982–3992.

COMPARATIVE ANALYSIS OF LEXICAL AND SEMANTIC SEARCH IN MULTILINGUAL MEDIA AGGREGATION WEB SERVICES

S.K. Simakin, L.M. Bozhukha

Abstract. *This paper presents the results of a comparative analysis of two search approaches in multilingual media content aggregation web services: lexical search based on the BM25 algorithm (Elasticsearch) and semantic search based on BGE-M3 vector embeddings (Qdrant). The MediaAggregator platform was developed to index 70,000 news articles in seven languages and provide a unified API for comparing search quality, latency, and response size. Experiments were conducted locally on a laptop with an AMD Ryzen 9 8945HS processor (Zen 4, 8 cores, 5.2 GHz boost) and 32 GB LPDDR5x RAM. Experimental results demonstrate that lexical search provides 2.7 times lower latency on average, while semantic search enables cross-lingual retrieval of relevant content regardless of query language, which is critical for multilingual web services.*

Keywords: *web service, search optimization, semantic search, Elasticsearch, vector database, BGE-M3, multilingual search, BM25, Qdrant.*