

## АНАЛІЗ ВИКОРИСТАННЯ МЕТОДІВ ПРИСКОРЕННЯ ВЕКТОРНОГО ПОШУКУ В РЕЛЯЦІЙНИХ СКБД

<sup>1</sup> Вінницький національний технічний університет;

**Анотація** У роботі проаналізовано методи прискорення векторного пошуку в реляційних системах керування базами даних. Розглянуто точний пошук за повним перебором, попереднє фільтрування кандидатів та наближений пошук найближчих сусідів на основі індексів IVF і HNSW. Основну увагу приділено особливостям реалізації векторного пошуку в PostgreSQL із розширенням pgvector та його узгодженню з атрибутивними SQL-предикатами.

**Ключові слова:** векторний пошук, реляційні СКБД, PostgreSQL, pgvector, ANN, HNSW, IVF, оптимізація запитів.

**Abstract** The paper analyzes methods for accelerating vector search in relational database management systems. Exact search based on full scan, candidate pre-filtering, and approximate nearest neighbor search using IVF and HNSW indexes are considered. Particular attention is paid to the implementation of vector search in PostgreSQL with the pgvector extension and to its integration with attribute-based SQL predicates.

**Keywords:** vector search, relational DBMS, PostgreSQL, pgvector, approximate nearest neighbor search, HNSW, IVF, query optimization.

### Вступ

Сучасні інформаційні системи дедалі ширше використовують векторні подання даних, або ембединги, для реалізації семантичного пошуку, рекомендаційних сервісів, Retrieval-Augmented Generation, пошуку схожих зображень та інших інтелектуальних функцій. У таких застосуваннях важливо не лише знайти найближчі вектори до заданого запиту, а й поєднати цей пошук із традиційними механізмами реляційної обробки даних, зокрема фільтрами за атрибутами, часовими обмеженнями, категоріями та бізнес-умовами. Це зумовлює актуальність дослідження векторного пошуку безпосередньо в реляційних системах керування базами даних [1].

Формально задача полягає у знаходженні  $k$  найближчих векторів  $x_i \in R^d$  до вектора запиту  $q$  за умови виконання предиката  $P(a_i)$ , визначеного на атрибутах запису. На рівні SQL це відповідає поєднанню умов *WHERE* з ранжуванням за векторною відстанню та оператором *LIMIT*. Однак точний пошук за повним перебором має складність  $O(Nd)$ , де  $N$  — кількість векторів, а  $d$  — їх розмірність, тому зі зростанням обсягу даних його використання стає обчислювально дорогим [2].

Метою роботи є аналіз використання методів прискорення векторного пошуку в реляційних СКБД та визначення умов ефективного застосування точного, наближеного і гібридного пошуку залежно від характеристик даних, параметрів запиту та вимог до продуктивності.

### Результати дослідження

Задачу пошуку найближчих векторів у реляційній СКБД можна подати у вигляді:

$$TopK(q) = \arg \min_{x_i: P(a_i)=true} dist(q, x_i), \quad (1)$$

де  $dist(q, x_i)$  метрикою відстані або мірою подібності між вектором запиту та вектором елемента бази даних. Така постановка відповідає SQL-запитам, у яких умови *WHERE* поєднуються з сортуванням за векторною близькістю та обмеженням кількості результатів.

Аналіз показує, що точний пошук є доцільним лише за невеликих обсягів даних або за високої селективності атрибутивних предикатів, коли попередня SQL-фільтрація суттєво зменшує множину кандидатів. У таких умовах реляційна СКБД може ефективно використовувати звичайні індекси за метаданими, після чого виконувати точне ранжування лише для обмеженої підмножини рядків. Проте для великих колекцій векторів практично необхідним стає використання наближеного пошуку найближчих сусідів [2-3].

У середовищі PostgreSQL із розширенням pgvector найбільш поширеними є індекси IVF та HNSW. Підхід IVF ґрунтується на попередньому розбитті векторного простору на кластери, що дає змогу під час виконання запиту аналізувати лише частину даних. Його перевагами є відносна простота побудови та можливість керувати компромісом між швидкістю і якістю пошуку. Водночас результативність IVF залежить від якості кластеризації та параметрів пошуку [4].

Індекс HNSW реалізує графову багаторівневу структуру навігації, яка зазвичай забезпечує вищі показники *recall@k* і меншу латентність для великих наборів високорозмірних векторів. Разом із тим цей підхід потребує більших витрат пам'яті та тривалішого часу побудови індексу. Отже, HNSW доцільно застосовувати в системах, орієнтованих на високу якість і швидкість читання, тоді як IVF є прийнятним компромісом у випадках, коли критичними є ресурсні обмеження та простіша організація індексної структури [5].

Принципово важливою для реляційних СКБД є не лише наявність ANN-індексу, а й спосіб його поєднання з SQL-предикатами. На практиці можливі дві базові схеми: спочатку виконується атрибутивна фільтрація, а потім пошук за векторною близькістю, або спочатку формується розширений список сусідів із ANN-індексу, після чого застосовується фільтрація за атрибутами. Ефективність кожної зі схем залежить від селективності предикатів, розподілу даних, параметрів індексу та здатності оптимізатора побудувати коректний план виконання запиту [6].

Отже, методи прискорення векторного пошуку в реляційних СКБД доцільно розглядати як комплексну систему, у якій визначальними є не лише індекси IVF або HNSW, а й узгодженість між векторною та атрибутивною індексацією. Найбільш ефективними в прикладних сценаріях є гібридні підходи, що забезпечують раціональний баланс між точністю, швидкістю та ресурсними витратами.

### Висновки

Отримані результати доцільно використовувати під час проектування систем семантичного пошуку, RAG-платформ, рекомендаційних сервісів та інших прикладних рішень, побудованих на PostgreSQL або інших реляційних СКБД із підтримкою векторних даних, оскільки вони дають змогу обґрунтовано вибирати між точним, наближеним і гібридним пошуком залежно від обсягу даних, селективності SQL-предикатів і вимог до продуктивності. Практично рекомендовано застосовувати точний пошук для невеликих наборів даних або високоселективних фільтрів, ANN-індекси — для великих векторних колекцій, а гібридні схеми — як найбільш результативний підхід для поєднання векторної та атрибутивної індексації в межах єдиного SQL-запиту. Перспективи подальших досліджень полягають у розробленні адаптивних методів вибору стратегії пошуку залежно від характеристик запиту, удосконаленні механізмів узгодження ANN-індексів із планувальником запитів реляційної СКБД, а також в оцінюванні ефективності таких підходів для різних типів ембедингів, навантажень і прикладних сценаріїв.

### СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Malkov Yu. A., Yashunin D. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. Vol. 42, no. 4. P. 824–836. DOI: 10.1109/TPAMI.2018.2889473.
2. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*. 2021. Vol. 7, no. 3. P. 535–547. DOI: 10.1109/TBDATA.2019.2921572.
3. Subramanya S. J., Devvrit, Simhadri H. V., Krishnaswamy R., Kadekodi R. DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node. In: *Advances in Neural Information Processing Systems (NeurIPS 2019)*. 2019.

4. PostgreSQL Global Development Group. PostgreSQL documentation [Electronic resource]. Available at: PostgreSQL official documentation.
5. Aumüller M., Bernhardsson E., Faithfull A. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. Information Systems. 2020. Vol. 87. Art. 101374. DOI: 10.1016/j.is.2019.02.006.
6. Muja M., Lowe D. G. Fast approximate nearest neighbors with automatic algorithm configuration. In: Proceedings of the Fourth International Conference on Computer Vision Theory and Applications (VISAPP 2009). Lisbon, Portugal, 2009. P. 331–340. DOI: 10.5220/0001787803310340.

**Кириленко Олександр Михайлович** - PhD, асистент кафедри комп'ютерних систем управління, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, e-mail: sasha.kyrylenko@gmail.com.

**Гармаш Володимир Володимирович** - к.т.н, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця.

**Kyrylenko Olexandr M.** - PhD, Assistant Professor of the Department of Computer Control Systems, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: sasha.kyrylenko@gmail.com.

**Garmash Volodymyr V.** — Candidate of Technical Sciences, Associate Professor of the Department of Automation and Intelligent Information Technologies,, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia