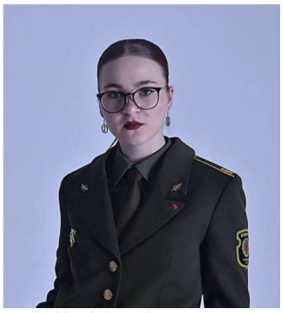
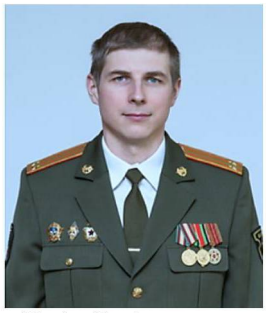
УДК 004.65

ПРИМЕНЕНИЕ ВЕКТОРНЫХ БАЗ ДАННЫХ ДЛЯ ОРГАНИЗАЦИИ ПОИСКА ТЕХНИЧЕСКОЙ И УЧЕБНОЙ ДОКУМЕНТАЦИИ В ОБЛАСТИ ТЕЛЕКОММУНИКАЦИЙ



А.В. Бардашевич
Курсант, БГУИР
sasha.bard9364@gmail.com

А.Ю. Савицкий
Старший преподаватель кафедры связи, БГУИР, кандидат военных наук
a.savitskij@bsuir.by



В.А. Федоренко
Начальник цикла кафедры связи, БГУИР
w.fedorenko@bsuir.by

А.В. Бардашевич
Курсант военного факультета Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с применением методов искусственного интеллекта и машинного обучения в образовательном процессе.

А.Ю. Савицкий
Окончил адъюнктуру Военной академии связи имени С.М. Буденного, кандидат военных наук. Область научных интересов связана с совершенствованием научно-методического аппарата оценки эффективности построения систем связи в общевойсковых соединениях и воинских частях, обоснованием принимаемых решений.

В.А. Федоренко
Окончил Военную академию Республики Беларусь. Область научных интересов связана с анализом существующих и разработкой новых методик защиты от несанкционированных атак транспортных сетей специального назначения.

Аннотация. Рассмотрена проблема неэффективного поиска учебной информации в подготовке специалистов в области телекоммуникаций, обусловленная семантической сложностью технической терминологии и межъязыковым барьером. Предложен подход к организации семантического поиска на основе векторных баз данных, обеспечивающий нахождение материалов по смыслу. Обоснована целесообразность использования векторного расширения для реляционной системы управления базами данных в образовательных учреждениях. Показано, что предложенное решение позволяет интегрировать векторный поиск в существующую инфраструктуру, обеспечивая поддержку межъязыковых запросов и фильтрацию по метаданным.

Ключевые слова: векторные базы данных, семантический поиск, pgvector, PostgreSQL, эмбеддинги, HNSW

Введение. Современная индустрия телекоммуникаций переживает период интенсивного технологического развития, обусловленный внедрением сетей пятого поколения, развитием технологий виртуализации сетевых функций, программно-конфигурируемых сетей и концепции Open RAN. По данным отчёта Ericsson Mobility Report [1], общий месячный глобальный трафик мобильных сетей достиг 200 эксабайт в четвертом квартале 2025 года, при этом годовой рост составил 22%.

Одновременно с усложнением телекоммуникационных систем экспоненциально растёт объем технической документации и образовательных материалов, что создает проблему информационной перегрузки в образовательном процессе. Традиционные методы поиска, по ключевым словам, демонстрируют низкую эффективность из-за семантической сложности предметной области: техническая терминология содержит много синонимов, аббревиатур и многозначных терминов, а значительная часть документации опубликована на английском языке, тогда как обучение ведется преимущественно на русском. В результате преподаватели и обучающиеся тратят значительное время на поиск релевантных материалов, а системы поиска по ключам не учитывают контекст запроса и не обеспечивают межъязыковое сопоставление информации.

Решением данной проблемы является применение векторных баз данных для организации семантического поиска учебного материала. Векторные системы управления базами данных (СУБД) позволяют осуществлять поиск по смыслу, а не по точному совпадению слов, что обеспечивает устойчивость к синонимии, поддержку межъязыковых запросов и учёт контекста использования терминов.

Анализ существующих решений. Традиционные подходы к поиску в образовательных ресурсах основываются на методах полнотекстового поиска (TF-IDF, BM25), которые оценивают релевантность документов по частоте встречаемости поисковых терминов [2]. Несмотря на вычислительную эффективность, эти методы не учитывают семантическую близость между словами, что критично для телекоммуникационной предметной области с ее высокой степенью синонимии и вариативности терминологии. Например, запрос «расчет запаса на замирания» не найдет документ с термином «fade margin calculation» из-за отсутствия лексического совпадения.

Развитие технологий обработки естественного языка привело к появлению методов семантического поиска на основе векторных представлений текста. Модели sentence embeddings, такие как Sentence-BERT [3], позволяют кодировать семантику текста в векторы фиксированной размерности, сохраняя возможность вычисления семантической близости через косинусное сходство. Для эффективного хранения и поиска по многомерным векторам применяются векторные базы данных, использующие алгоритмы приближенного поиска ближайших соседей, в частности Hierarchical Navigable Small World (HNSW) – один из самых быстрых и точных алгоритмов для поиска приблизительных ближайших соседей в многомерных векторных пространствах, обеспечивающий логарифмическую сложность поиска O(log n) [4].

Современный рынок предлагает различные решения для векторного поиска: специализированные СУБД (Qdrant, Weaviate, Pinecone), библиотеки и расширения для традиционных реляционных систем. Для образовательных учреждений телекоммуникационного профиля наиболее перспективным представляется использование расширения pgvector для СУБД PostgreSQL.

Данный подход обеспечивает ряд преимуществ: возможность выполнения сложных SQL-запросов с фильтрацией по метаданным (домен, технология, уровень сложности), ACID-совместимость и надежность, характерные для PostgreSQL, а также упрощение развертывания за счёт использования единой инфраструктуры для хранения как структурированных данных, так и векторных эмбеддингов. Исследования показывают, что производительность pgvector при работе с количеством до 1 миллиона векторов сопоставима со специализированными решениями, при этом время отклика составляет менее 100 мс. для большинства запросов.

Архитектура Retrieval-Augmented Generation сочетает извлечение информации из векторной базы данных с генеративными возможностями языковых моделей, что позволяет не только находить релевантные учебные материалы, но и формировать связные ответы с цитированием источников [5]. В контексте образовательных систем телекоммуникационного профиля данная архитектура обеспечивает достоверность

предоставляемой информации за счёт опоры на верифицированную предметную базу данных. Retrieval-компонент архитектуры реализован на основе расширения pgvector для PostgreSQL, что обеспечивает эффективный семантический поиск с фильтрацией по метаданным (домен связи, технология, уровень сложности). Такой подход позволяет интегрировать систему в существующую инфраструктуру образовательных учреждений без развертывания дополнительных специализированных решений, сохраняя при этом возможность выполнения сложных запросов и ACID-совместимость.

Архитектура предлагаемой системы семантического поиска (рисунок 1) построена на основе расширения pgvector для СУБД PostgreSQL и включает четыре основных компонента: модуль сбора и предобработки документов, модуль векторизации текста, векторное хранилище на базе pgvector и модуль векторного поиска.

Рисунок 1. Архитектура семантического поиска

Обработка данных начинается с подготовки исходных документов, включающая очистку текста, стандартизацию терминов и сегментацию на логические фрагменты. Полученные блоки кодируются в векторную форму с помощью модели эмбеддингов и загружаются в СУБД PostgreSQL, где к ним добавляются описательные метаданные (домен связи, технология, сложность, источник). Поиск реализуется через векторизацию пользовательского запроса аналогичным методом и сопоставление с индексированными данными посредством алгоритма HNSW. Ранжирование результатов выполняется согласно

метрике косинусного сходства, обеспечивая выдачу релевантных материалов с информацией об источнике.

Процесс векторизации учебных материалов начинается с извлечения текста из документов различных форматов (DOCX, HTML, PDF). Для очистки текста применяется нормализация Unicode-символов, удаление избыточных пробелов и переносов строк, сохранение технических обозначений (частоты, мощности, формулы). Особое внимание уделяется обработке телекоммуникационной терминологии: аббревиатуры раскрываются при первом упоминании, синонимы приводятся к канонической форме, английские термины сохраняются в оригинале для обеспечения межъязыкового поиска. Разбиение текста на чанки выполняется рекурсивным методом с приоритетом сохранения логической структуры документа (разделы, подразделы, абзацы). На основе анализа структуры технической документации обоснованы параметры чанкинга: размер чанков 800 токенов с перекрытием 150 токенов, что обеспечивает баланс между сохранением контекста и точностью поиска (таблица 1).

Таблица 1. Параметры чанкинга

Размер чанка (токены)	Перекрытие (токены)	Точность поиска	Время индексации	Рекомендации
512	100	0.82	Быстро	Для простых документов
800	150	0.89	Средне	Оптимально
1024	200	0.88	Медленно	Для сложных текстов

Для преобразования текста в векторные представления выбрана модель paraphrase-multilingual-MiniLM-L12-v2, поддерживающая русский и английский языки. Данная модель обеспечивает размерность вектора 384, что представляет собой компромисс между качеством эмбеддингов и требованиями к памяти. Модель обучена на разных наборах данных, включая техническую литературу, что обеспечивает качественное кодирование телекоммуникационной терминологии. Векторизация выполняется на стороне приложения с использованием библиотеки sentence-transformers, полученные векторы передаются в PostgreSQL для хранения в таблице с расширением pgvector.

Хранение векторных эмбеддингов организовано в таблице PostgreSQL со следующей структурой: уникальный идентификатор чанка, текстовое содержимое, вектор, метаданные в формате JSONB (домен, поддомен, технология или стандарт, тема, уровень сложности материала, тип контента, источник, страница). Для ускорения поиска по векторам создается индекс HNSW с параметрами: метрика косинусного сходства, количество связей каждого узла в графе M = 16, размер списка кандидатов при построении ef_construction = 128 – задает качество построения индекса, при поиске ef_search = 100 – контролирует точность поиска. Данные параметры обеспечивают время отклика менее 100 мс. при размере базы до 100 000 векторов. Преимуществом использования PostgreSQL является возможность выполнения гибридных запросов, сочетающих векторный поиск с фильтрацией по метаданным через стандартные SQL-условия WHERE.

Алгоритм семантического поиска включает следующие этапы (рисунок 2): прием пользовательского запроса, предобработка текста (очистка, нормализация), векторизация запроса той же моделью эмбеддингов, что и для документов, выполнение SQL-запроса к базе данных с поиском ближайших соседей и фильтрацией по метаданным, ранжирование результатов по косинусному сходству, возврат наиболее релевантных результатов с метаданными [6].

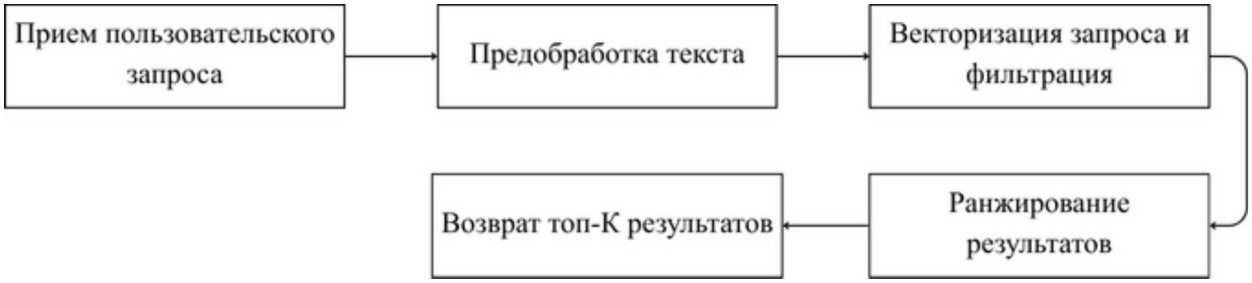


Рисунок 2. Алгоритм семантического поиска

Для повышения производительности реализовано кеширование частых запросов с использованием хеша запроса в качестве ключа. Поддерживается гибридный поиск с возможностью комбинации семантического и полнотекстового поиска через встроенные средства PostgreSQL (tsvector/tsquery), что позволяет дополнительно повысить точность для запросов с точными техническими терминами.

Для обеспечения актуальности базы данных разработан процесс инкрементального обновления: новые документы проходят полный цикл обработки и добавляются в базу без необходимости переиндексации существующих данных [7]. Версионирование стандартов и спецификаций поддерживается через метаданные с указанием версии документа и даты публикации. Система предоставляет API для интеграции с внешними образовательными платформами и веб-интерфейсом, что обеспечивает возможность использования как в учебном процессе ВУЗов, так и в корпоративных системах обучения инженеров по телекоммуникациям.

Заключение. В работе рассмотрена задача организации семантического поиска технической и учебной документации при подготовке специалистов в области телекоммуникаций. Предложенный подход на основе векторных баз данных обеспечивает семантический поиск, который находит материалы по смыслу, а не по лексическому совпадению, что устраняет проблемы синонимии терминов и межъязыкового барьера.

Использование расширения pgvector для PostgreSQL позволяет интегрировать векторный поиск в существующую инфраструктуру образовательных учреждений без развертывания дополнительных специализированных СУБД, сохраняя при этом возможность сложных запросов с фильтрацией по метаданным.

Рассмотренная архитектура векторизации технической и учебной документации позволит сократить время поиска учебной информации, повысить эффективность самостоятельной работы обучающихся может быть применена при создании образовательных ресурсов для других инженерных направлений со схожей спецификой предметной области.

Список литературы

[1] Ericsson Mobility Report: November 2025 / Ericsson. – Stockholm: Ericsson AB, 2025. – URL: https://www.ericsson.com/en/reports-and-papers/mobility-report (дата обращения: 01.03.2026).

[2] Маннинг, К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – М.: Вильямс, 2019. – 528 с.

[3] Платонов, А. В. Машинное обучение / А. В. Платонов. – М.: Юрайт, 2021. – 267 с.

[4] Литвиненко, В. И. Обработка естественного языка / В. И. Литвиненко. – М.: Юрайт, 2024. – 245 с.

[5] Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus [et al.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 9459–9474.

[6] Моргунов, Е. П. PostgreSQL. Основы языка SQL / Е. П. Моргунов. – СПб.: БХВ-Петербург, 2019. – 304 с.

[7] Кузнецов, С. Д. Основы баз данных / С. Д. Кузнецов. – М.: Интуит, 2022. – 488 с.

Авторский вклад

Бардашевич Александра Валерьевна – разработка алгоритма семантического поиска с использованием индекса HNSW, проектирование процесса предобработки документов и нормализации телекоммуникационной терминологии, обеспечение межъязыковой поддержки поиска.

Савицкий Алексей Юрьевич – постановка задачи исследования, разработка архитектуры системы семантического поиска, обоснование выбора расширения pgvector для PostgreSQL, подготовка иллюстрационного материала.

Федоренко Владимир Александрович – разработка методологии векторизации технической документации, обоснование параметров чанкинга и выбора модели эмбеддингов, проектирование структуры метаданных, анализ существующих решений векторного поиска

APPLICATION OF VECTOR DATABASES FOR STORAGE AND SEARCH OF EDUCATIONAL MATERIALS IN THE FIELD OF TELECOMMUNICATIONS

A.V. Bardashevich
Cadet, BSUIR

A.Yu. Savitsky
Senior Lecturer of the Department of Communications, BSUIR, Ph.D. in Military Sciences

V.A. Fedorenko
Head of the Department of Communications, BSUIR

Abstract. This article examines the problem of inefficient search for educational information in the training of telecommunications specialists, caused by the semantic complexity of technical terminology and the interlingual barrier. An approach to organizing semantic search based on vector databases is proposed, ensuring the search for materials by meaning. The feasibility of using a vector extension for a relational database management system in educational institutions is substantiated. It is shown that the proposed solution enables the integration of vector search into the existing infrastructure, providing support for cross-lingual queries and metadata filtering.

Keywords: vector databases, semantic search, pgvector, PostgreSQL, embeddings, HNSW