УДК 004.65

ОБЗОР ВАРИАНТОВ ИСПОЛЬЗОВАНИЯ ВЕКТОРНЫХ СИСТЕМ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ

Загальский Е.В. (ИТМО)

Научный руководитель – доктор технических наук, профессор Бессмертный И.А. (ИТМО)

Введение. В настоящее время векторные системы управления базами данных (СУБД) занимают важное место в мире разработки приложений искусственного интеллекта и являются значимым компонентом современного подхода к управлению данными. Развитие и применение векторных СУБД было обусловлено необходимостью в эффективной обработке и хранении векторных представлений (эмбеддингов), которые получаются в результате преобразования естественно языковых текстов, изображений, аудио и видео. Векторные СУБД применяются в приложениях рекомендательных систем, обратного поиска изображений, диалоговых систем и электронной коммерции [1].

Основная часть. В эпоху развития приложений искусственного интеллекта, в частности генеративных нейронных сетей, становятся актуальными исследования в области векторных баз данных для решения задач эффективной обработки и анализа данных. Исходя из особенностей современных приложений искусственного интеллекта, которые генерируют векторные представления - эмбеддинги, появилась необходимость в эффективном управлении и взаимодействии с данными подобного типа.

Эмбеддинг - числовой вектор, который получается в результате преобразования текста, изображений, документов и иных видов данных. Например, двумерным числовым вектором является (1.7, 2.5), а четырехмерным (0, 1.2, 2.4, 3.2). Таким образом, числовой вектор возможно использовать в качестве многомерных координат для измерения сходства и производить математические операции. Следовательно, эмбеддинги позволяют выполнять векторный поиск по сходству и семантический поиск.

Векторная система управления базами данных — новый тип СУБД, ориентированный на эффективное управление многомерными векторными представлениями данных. Стоит отметить, что современные векторные базы данных хранят векторные представления вместе с исходными данными, например Milvus (векторная СУБД с открытым исходным кодом), что обеспечивает возможность использования как векторного, так и классического поиска по ключевым словам. В отличие от реляционных баз данных, где запросы принимают такие формулировки, как «найти определенного пользователя» или «найти товары с определенным статусом», векторные запросы имеют вид «найти k наиболее похожих изображений на определенное изображение» или «найти наиболее подходящие рестораны, учитывая мою профессию» [1]. Были выделены следующие варианты использования векторных СУБД:

1) Использование векторных баз данных для семантического поиска и хранения векторных представлений для различных модальностей [1];
2) Использование векторных баз данных для поиска по сходству [1];
3) Применение векторных баз данных в качестве базы знаний для повышения эффективности метода Retrieval Augmented Generation (RAG) [2];
4) Использование векторных баз данных в качестве слоя кэширования при взаимодействии с большими языковыми моделями через API [3].

Выводы. Проведен обзор вариантов использования векторных баз данных для задач векторного поиска по сходству, семантического поиска и хранения векторных представлений. Также рассмотрены подходы для повышения эффективности метода RAG и кэширования при взаимодействии с большими языковыми моделями с помощью векторных СУБД.

Список использованных источников:

1. Taipalus, T. (2023). Vector database management systems: Fundamental concepts, use-cases, and current challenges. arXiv preprint arXiv:2309.11322.
2. Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. arXiv preprint arXiv:2310.11703.
3. Jing, Z., Su, Y., Han, Y., Yuan, B., Liu, C., Xu, H., & Chen, K. (2024). When Large Language Models Meet Vector Databases: A Survey. arXiv preprint arXiv:2402.01763.