УДК 004.8

Сравнительный анализ алгоритмов индексации векторных представлений данных

Загальский Е.В. (ИТМО)

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

Введение. Методы приближенного поиска ближайших соседей (ANN) являются значимым компонентом интеллектуального анализа данных, компьютерного зрения, обработки естественного языка, векторных баз данных и рекомендательных систем. Стоит отметить, что модели глубокого обучения обладают способностью преобразовывать естественно-языковый текст, изображения, аудио и видео в векторные представления данных. Например, в современных системах генерации с дополненой выборкой (Retrieval augmented generation, RAG) применяются методы ANN для поиска информации и последующей интеграции в процесс генерации ответа больших языковых моделей (LLM) для повышения точности и актуальности результирующего вывода. В данном контексте, индексация имеет решающее значение, поскольку значительно повышает эффективность и точность поиска при работе с большими наборами данных в приложениях искусственного интеллекта [1].

Основная часть. В настоящее время существует множество алгоритмов индексации векторных представлений данных, которые возможно разделить по их типу, а именно [1]:

1) индекс на основе древовидной структуры (Tree-based index);
2) индекс на основе хеширования (Hash-based index);
3) графовый индекс (Graph-based index);
4) индекс на основе квантования (Quantization-based index [1].

В качестве примера, индексом на основе древовидной структуры является ANNOY (Approximate Nearest Neighbors Oh Yeah). Для индексации векторных представлений данных на основе хеширования применяется вероятностный метод понижения размерности многомерных данных (Locality sensitive hashing, LSH). Иерархический маленький мир (Hierarchical Navigable Small Worlds, HNSW), DiskANN, FreshDiskANN и их модификации являются экземплярами графовых индексов. Примером индекса на основе квантования является Inverted file index (IVF). Также к индексам на основе квантования относятся индексы с применением следующих методов сжатия: квантование по продукту (Product Quantization, PQ), скалярное квантование (Scalar Quantization, SQ) и бинарное квантование (Binary Quantization, BQ). Таким образом к индексам на основе квантования относятся HNSW+PQ, IVF+PQ и другие комбинации. Стоит отметить, что вышеперечисленные индексы возможно разделить по типу хранилища индекса, а именно:

1) в оперативной памяти (RAM) – ANNOY, HNSW, IVF;
2) в твердотельном накопителе (SSD) – DiskANN, FreshDiskANN;
3) в графическом процессоре (GPU) – CAGRA [2].

Среди выбранных алгоритмов, графовые индексы демонстрируют самые высокие показатели скорости и точности (Recall) поиска по сравнению с другими индексами, но имеют значительные ограничения для масштабирования и потокового векторного поиска (streaming vector search), включая последовательные операции вставки, обновления и удаления данных, что является важным аспектом современных приложений искусственного интеллекта, векторных баз данных и систем с использованием метода RAG [3]. Также графовые индексы потребляют значительное количество вычислительных ресурсов (RAM – HNSW, SSD – DiskANN, GPU – CAGRA) и переиндексация больших наборов данных требует значительное количество времени.

Выводы. Проведен сравнительный анализ актуальных алгоритмов индексации векторных представлений данных. Проведено экспериментальное сравнение алгоритмов индексации. Выявлены актуальные научные проблемы и выбрано дальнейшее направление

исследований, а именно проведение экспериментов с модификациями алгоритмов HNSW, DiskANN и IVF.

Список использованных источников:

1. Xiao, W., Zhan, Y., Xi, R., Hou, M., & Liao, J. (2024). Enhancing HNSW Index for Real-Time Updates: Addressing Unreachable Points and Performance Degradation. arXiv preprint arXiv:2407.07871.

2. Ootomo, H., Naruse, A., Nolet, C., Wang, R., Feher, T., & Wang, Y. (2024, May). Cagra: Highly parallel graph construction and approximate nearest neighbor search for gpus. In 2024 IEEE 40th International Conference on Data Engineering (ICDE) (pp. 4236-4247). IEEE.

3. K. Lu, M. Kudo, C. Xiao, and Y. Ishikawa, “Hvs: hierarchical graph structure based on voronoi diagrams for solving approximate nearest neighbor search,” Proc. VLDB Endow., vol. 15, no. 2, p. 246–258, oct 2021. [Online]. Available: https://doi.org/10.14778/3489496.3489506