Применимость линейных методов снижения размерности для приближенного поиска ближайших соседей на запросах вне исходного распределения

П. А. Алисейчик*, Е. А. Снегова‡

Исследуется применение линейных методов снижения размерности и квантования для ускорения приближённого поиска ближайших соседей (ANNS) при различии распределений индексируемых векторов и запросов. Показано, что отбрасывание компонент, малых для индексируемых векторов, может приводить к существенной ошибке скалярного произведения на запросах вне исходного распределения (OOD). Предлагается последовательное разложение пространства на ортогональные подпространства, позволяющее учитывать компоненты, значимые для OOD-запросов. На базе Text-to-Image размерности 200 предложенный подход сокращает объём хранения векторов в 1,67 раза и ускоряет поиск в 1,14 раза по сравнению с PCA-квантованием.

Ключевые слова: приближённый поиск ближайших соседей, снижение размерности, запросы вне исходного распределения, линейные преобразования, квантование.

1. Введение

Приближённый поиск ближайших соседей (Approximate Nearest Neighbor Search, ANNS) широко применяется в векторных поисковых системах и мультимодальных приложениях. Для ускорения поиска используются методы снижения размерности, позволяющие сократить объём хранимых данных и вычислительную сложность. К ним относятся метод главных компонент (Principal Component Analysis, PCA) и другие линейные преобразования; обзор и сравнительная оценка таких методов представлены в работе [1]. Ограничения методов оптимизации операций сравнения расстояний (Distance Comparison Operations, DCO) исследуются в [2].

Алисейчик Павел Александрович — к. ф.-м. н., в. н. с. кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: paa.idea@mail.ru.

Aliseychik Pavel A. — Candidate of Physical and Mathematical Sciences, Leading Researcher, Department of Mathematical Theory of Intelligent Systems, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University.

‡*Снегова Елена Александровна* — к. ф.-м. н., e-mail: lenasnegova@gmail.com.

Snegova Elena A. — Candidate of Physical and Mathematical Sciences.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

Важной проблемой является различие распределений индексируемых векторов и запросов (Out-of-Distribution, OOD). В методе LeanVec [3] для их обработки применяются отдельные линейные проекции запросов и ключей.

2. Постановка задачи и анализ линейного снижения размерности

Рассматривается поиск k ближайших векторов из базы $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^D$, для запроса $y \in \mathbb{R}^D$. В данной работе рассматриваются поиск по скалярному произведению (IP) и евклидову расстоянию (L2).

Как правило, рассматриваются центрированные векторы. Пусть x_0 и y_0 — математические ожидания распределений ключей и запросов. После представления $x = x_0 + x_1$, $y = y_0 + y_1$ получаем

$$x^T y = x_1^T y_1 + x_1^T y_0 + x_0^T y. \tag{1}$$

Слагаемое $x_1^T y_0$ предварительно вычисляется для ключей, а $x_0^T y$ — один раз для запроса. Далее рассматриваются центрированные векторы x_1 и y_1, обозначаемые x и y.

Снижение размерности обычно строится по распределению индексируемых векторов, например с помощью PCA [1] и обучаемых преобразований [3]. Рассмотрим ортогональное разложение $V = V_2 \oplus V_3$, где V_2 соответствует значимым для ключей координатам:

$$x = P_2(x) + P_3(x). \tag{2}$$

Если $\|P_3(x)\|_2$ мала, соответствующие координаты можно отбросить. Однако для OOD-запросов $\|P_3(y)\|_2$ может быть значительной. В силу ортогональности

$$x^T y = P_2(x)^T P_2(y) + P_3(x)^T P_3(y). \tag{3}$$

Поэтому малая проекция ключа не гарантирует малой ошибки скалярного произведения.

Для учёта компонент, значимых для запросов, повторно разделим $V_3 = V_4 \oplus V_5$, выбрав V_5 так, чтобы проекции запросов на него имели малую норму. Тогда

$$P_3(x)^T P_3(y) = P_4(x)^T P_4(y) + P_5(x)^T P_5(y), \tag{4}$$

откуда

$$x^T y = P_2(x)^T P_2(y) + P_4(x)^T P_4(y) + P_5(x)^T P_5(y). \tag{5}$$

При малой $\|P_3(x)\|_2$ последнее слагаемое можно отбросить, если его вклад для запросов также невелик. Ошибка зависит от норм проекций и их взаимной ориентации.

Квантование позволяет дополнительно сократить объём хранения и ускорить вычисление скалярных произведений за счёт уменьшения числа бит на координату.

Эксперименты проведены на Yandex Text-to-Image (10^6 векторов размерности 200 и 10^5 текстовых запросов из стандартного публичного набора) с использованием графа Vamana и оптимизаций для OOD.

Таблица 1: Результаты поиска на Text-to-Image (IP, recall@10 $\approx 0{,}95$)

Способ квантования	Размер вектора, байт	Запросов в секунду
Без квантования, float32	800	5500
PCA, uint8	200	12300
Предлагаемый метод: 40 значений uint8 и 160 значений по 4 бита	120	14000

При сопоставимом качестве поиска (recall@10 $\approx 0{,}95$) предложенный метод по сравнению с PCA-квантованием сокращает объём хранения векторов в 1,67 раза и ускоряет поиск в 1,14 раза.

Дополнительно проведены эксперименты на базе SIFT-1M (10^6 векторов размерности 128, метрика L2). Для моделирования OOD случайный поворот и смещение применялись к запросам. Поиск также выполнялся по графу Vamana с оптимизациями для OOD.

Таблица 2: Результаты поиска на базе SIFT-1M (L2, OOD, recall $\approx 0{,}95$)

Метод	Размер вектора, байт	Запросов в секунду
float32, полная размерность	512	~ 8500
PCA, dim=64	260	~ 3800
Предлагаемый метод, dim=64	260	~ 12400

Список литературы

[1] Z. Wang, H. Xiong, Q. Wang, Z. He, P. Wang, T. Palpanas, W. Wang, “Dimensionality-Reduction Techniques for Approximate Nearest Neighbor Search: A Survey and Evaluation”, *IEEE Data Engineering Bulletin*, **48**:3 (2024), 63–80.

[2] Z. Zheng, Y. Zeng, C. Liu, Y. Chi, B. Yang, Y. Tong, “Distance Comparison Operations Are Not Silver Bullets in Vector Similarity Search: A Benchmark Study on Their Merits and Limits”, *Proceedings of the IEEE 42nd International*

Conference on Data Engineering, 2026, 2063–2076. DOI: 10.1109/ICDE65706.2026.00156.

[3] M. Tepper, I. S. Bhati, C. Aguerrebere, M. Hildebrand, T. L. Willke, "LeanVec: Searching Vectors Faster by Making Them Fit", *Transactions on Machine Learning Research*, **2024** (2024).

Статья поступила 9 ноября 2026 г.

Applicability of Linear Dimensionality Reduction Methods to Approximate Nearest Neighbor Search with Out-of-Distribution Queries

P. A. Aliseychik, E. A. Snegova

We study linear dimensionality reduction and quantization for approximate nearest neighbor search (ANNS) when the distributions of indexed vectors and queries differ. We show that discarding components that are small for indexed vectors may cause significant dot-product errors for out-of-distribution (OOD) queries. We propose a sequential decomposition into orthogonal subspaces to account for components important to OOD queries. On the 200-dimensional Text-to-Image dataset, the proposed approach reduces vector storage by a factor of 1.67 and accelerates search by a factor of 1.14 compared with PCA-based quantization.

Keywords: approximate nearest neighbor search, dimensionality reduction, out-of-distribution queries, linear transformations, quantization.

References

[1] Z. Wang, H. Xiong, Q. Wang, Z. He, P. Wang, T. Palpanas, W. Wang, "Dimensionality-Reduction Techniques for Approximate Nearest Neighbor Search: A Survey and Evaluation", *IEEE Data Engineering Bulletin*, **48**:3 (2024), 63–80.

[2] Z. Zheng, Y. Zeng, C. Liu, Y. Chi, B. Yang, Y. Tong, "Distance Comparison Operations Are Not Silver Bullets in Vector Similarity Search: A Benchmark Study on Their Merits and Limits", *Proceedings of the IEEE 42nd International Conference on Data Engineering*, 2026, 2063–2076. DOI: 10.1109/ICDE65706.2026.00156.

[3] M. Tepper, I. S. Bhati, C. Aguerrebere, M. Hildebrand, T. L. Willke, "LeanVec: Searching Vectors Faster by Making Them Fit", *Transactions on Machine Learning Research*, **2024** (2024).

Received on November 9, 2026