На правах рукописи

Казаковцев Владимир Львович

АЛГОРИТМЫ УСКОРЕННОГО ПОИСКА В ВЕКТОРНЫХ БАЗАХ ДАННЫХ

2.3.1 - Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2026

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Сибирский федеральный университет», г. Красноярск.

Научный руководитель: доктор технических наук, профессор
Ступина Алена Александровна

Официальные оппоненты: **Пимонов Александр Григорьевич,**
доктор технических наук, профессор,
ФГБОУ ВО «Кузбасский государственный технический университет имени Т. Ф. Горбачева» (КузГТУ), г. Кемерово,
заведующий кафедрой прикладных информационных технологий
Царев Роман Юрьевич,
кандидат технических наук, доцент,
ФГБОУ ВО «МИРЭА – Российский технологический университет» (РТУ МИРЭА), г. Москва, доцент кафедры высшей математики института искусственного интеллекта

Ведущая организация: федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный технический университет»

Защита состоится 19 июня 2026 года в 15:00 часов на заседании диссертационного совета 24.2.403.01, созданного на базе ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» по адресу: 660037, г. Красноярск, пр. им. газеты «Красноярский рабочий» 31, зал заседаний диссертационного совета, ауд. Л-205

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» и на сайте https://www.sibsau.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью, просим отправлять по адресу: 660037, Россия, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» (СибГУ им. М.Ф. Решетнева), Диссертационный совет E-mail: dissovet@sibsau.ru

Автореферат разослан «___» _____ 2026 г.

Ученый секретарь
диссертационного совета Панфилов Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Задачи приближенного поиска ближайших соседей получили свое развитие в связи с широким распространением методов представления данных в виде высокоразмерных векторных признаков (наборов числовых характеристик объектов) и стремительным ростом объемов информации, обрабатываемой в современных информационных и интеллектуальных системах. Во многих прикладных областях, таких как семантический поиск, рекомендательные сервисы, анализ изображений и аудиосигналов, обнаружение аномалий и так далее, требуется многократное выполнение запросов на поиск близких объектов по метрике, однако точные методы становятся вычислительно затратными при большом количестве объектов в наборе данных и высокой размерности пространств признаков. Приближенные алгоритмы позволяют существенно сократить время ответа и потребление памяти при контролируемом снижении качества результатов, обеспечивая практическую реализуемость решений в условиях ограничений по задержке и стоимости вычислений.

Некоторые алгоритмы приближенного поиска ближайших соседей используют методы автоматической группировки данных, также называемые алгоритмами кластеризации. Эти алгоритмы решают задачу кластеризации, которая не имеет строгого формального определения, но неформальным образом может быть описана как задача разделения множества объектов на подмножества таким образом, чтобы объекты в одном подмножестве (кластере) были как можно сильнее схожи друг с другом и как можно сильнее отличались от объектов в других подмножества. Алгоритмы кластеризации входят в число базовых методов машинного обучения и анализа данных, которые находят широкое применение в различных областях, включая обработку результатов научных экспериментов, интернет-технологий, в частности обработку данных социальных сетей, обработку медицинских данных, финансы и другие. Рост объема и размерности собираемых и сохраняемых данных требует специальных методов и моделей для их эффективной обработки. Кластеризация упрощает анализ больших и сложных для ручной обработки данных, структурируя их и позволяя исследователям и практикам сосредоточиться на значимых группах объектов, а также позволяет находить закономерности в данных и выявлять взаимосвязи между объектами.

Одним из методов приближенного поиска ближайших соседей, использующим кластеризацию, является инвертированный индекс (англ. Inverted File Index, IVF). Метод инвертированного индекса заключается в том, что все объекты в базе данных разбиваются на кластеры, и каждый кластер характеризуется его центроидом, которому в соответствие ставится список всех объектов внутри кластера. При поиске ближайших к вектору-запросу объектов базы данных на первом шаге выбирается несколько ближайших к вектору-запросу центроидов, а сам поиск ближайших соседей проводится не по всей базе данных, а внутри кластеров, определяемых этими центроидами, что позволяет значительно сократить количество необходимых для обработки запроса вычислительных ресурсов. Приближенный поиск ближайших соседей играет важную роль в современных информационных системах, таких как рекомен-

дательные сервисы, компьютерное зрение и анализ данных. В условиях постоянно растущих объемов данных и задач поиска по сложным многомерным структурам, методы точного поиска не могут обеспечить приемлемую скорость поиска. Актуальность методов приближенного поиска обуславливается необходимостью быстрой обработки больших массивов информации, которые позволяют ускорить время отклика, сохраняя приемлемую точность поиска.

Развитие носителей информации и интернет-технологий ставит новые задачи, в том числе задачу обработки мультимодальных данных большого объема. Мультимодальными данными называют данные, представленные одновременно в двух и более модальностях (разнородных видах представления информации), например текст, изображение, аудио или видео, относящиеся к одному и тому же объекту или событию.

Изложенное выше свидетельствует о перспективности развития алгоритмов автоматической группировки данных для решения широкого круга задач, включая разработку методов кластеризации мультимодальных данных и алгоритмов приближенного поиска ближайших соседей.

Степень научной разработанности проблемы. Среди важнейших алгоритмов поиска ближайших соседей можно выделить предложенный Ю. Мальковым и Д. Яшуниным HNSW (англ. Hierarchical Navigable Small World), который использует многослойную иерархическую структуру графов. Другим наиболее распространенным алгоритмом является алгоритм поиска по обратному индексу (англ. Inverted File Index, IVF), который для ограничения области поиска использует кластерную структуру данных. Сложно выделить одного или нескольких авторов этого алгоритма, однако важный вклад в развитие этого подхода внесли такие ученые, как Л. Ванг, А. Бабенко, С. Брух и другие.

Самой распространенной моделью кластеризации является модель *k*-средних, предложенная в 1957 году Г. Штейнгаузом. С. Ллойд позднее предложил алгоритм, который известен как алгоритм Ллойда или процедура *k*-средних. Значительный вклад в развитие теории кластеризации внесли Ц. Дрезнер, О. Алп, Э. Эркут и Н. Младенович, считающиеся сегодня классиками в этой области. А.Н. Антамошкин и другие отечественные ученые развили теорию эвристических алгоритмов кластеризации. Также важно отметить Л. Кауфмана и П. Руссива, представивших модель *k*-медоид, а также Х. Хамахера, П. Хансена и Ю. Кочетова, которые развивали метод чередующихся окрестностей (Variable Neighbourhood Search). В.В. Шаламовым, В.А. Ефимовой, С.Б. Муравьевым и А.Ф. Фильченковым был разработан алгоритм для автоматического выбора алгоритма кластеризации и настройки его гиперпараметров MASSCAH (Multi-armed simultanous selection of clustering algorithm and its hyperparameters).

Красноярской научной школой: А.Н. Антамошкиным, Л.А. Казаковцевым, А.А. Ступиной, Г.Ш. Шкабериной и др. был внесен значительный вклад в развитие эволюционных алгоритмов кластеризации, жадных агломеративных эвристических процедур, различных операторов мутации. В.И. Голованом, В.В. Орловым и Л.А. Ка-

заковцевым были развиты методы выделения однородных партий электрорадиоизделий для систем с повышенными требованиями к качеству, однородности и отказоустойчивости микросхем.

К. Ли, М. Чжаном, Д.Г. Андерсеном и Ю. Хе предложен алгоритм адаптивного поиска приближенных ближайших соседей для методов HNSW и IVF. Авторы рассматривают расстояния между объектами и центрами кластеров и исходя из этих данных их алгоритм определяет глубину поиска.

Обработка мультимодальных данных является быстро развивающейся областью, в которой сложно выделить основные, фундаментальные работы. Ч. Чен и др. предложили метод агрегации модальностей на основе энкодер-декодерной архитектуры, использующий для агрегации модальностей метод глубокого обучения, то есть многослойные нейронные сети, требующие обучения на размеченных данных и настройки параметров обучения. Другие подходы, также использующие глубокое обучение, были предложены в работах М. Аль Рахала, В. Ванга и других. В. Ванг с соавторами показал достаточно эффективную модель автоэнкодера для объединения нескольких модальностей в единое векторное представление (эмбеддинг), однако, такая модель также требует обучения или дообучения на размеченных данных, а также не может быть эффективно реализована на вычислительном кластере. Н. Шривастава и Р.Р. Салахутдинов предложили для создания единого векторного представления использовать две машины Больцмана, то есть две рекуррентных нейронных сети, которые также требуют обучения на размеченных данных и больших вычислительных ресурсов. Метод ADAPT, разработанный Г.С. Лбовым и соавторами в 1980-х гг., является принципиально иным подходом: все пространство поиска представляется как конечное множество, где каждая переменная может принимать конечный набор значений.

Несмотря на обширные накопленные знания, методики и технологии, в существующих методах кластерного анализа есть недостатки. Так, современные задачи требуют агрегации разных типов данных, так как информация об одиночном объекте может быть представлена более чем одним вектором. Также следует заметить, что в области эволюционных алгоритмов есть поле для исследований: могут быть разработаны более точные и стабильные алгоритмы, а также более специализированные алгоритмы кластеризации на основе эволюционных алгоритмов.

Целью настоящей работы является повышение эффективности алгоритмов приближенного поиска ближайших соседей.

Для достижения цели были поставлены следующие **задачи:**

1. Провести анализ проблем, возникающих при применении методов приближенного поиска ближайших соседей, а также анализ методов, позволяющих повысить быстродействие алгоритмов приближенного поиска ближайших соседей без потери точности.

2. Разработать алгоритм классификации запросов приближенного поиска ближайших соседей с использованием IVF-индекса, основанный на оценке доли

эффективных (результативных) кластеров на начальных этапах поиска, позволяющий определить сложность запроса и предсказать количество кластеров, в которых находятся ближайшие соседи.

3. Разработать адаптивный алгоритм поиска данных в векторной базе данных на основе IVF-индекса с использованием классификатора сложности запросов по результатам предварительного поиска, ускоряющий процесс поиска ближайших соседей без потери качества.

4. Разработать алгоритмы на основе жадной агломеративной процедуры, которые бы представляли компромисс между достигаемым значением целевой функции и необходимыми для выполнения вычислительными и временными ресурсами, а также работали стабильно и конкурентоспособно по сравнению с другими алгоритмами (по внешним и внутренним мерам качества кластеризации).

5. Разработать меру расстояния, агрегирующую расстояния между объектами, рассчитанные по отдельным модальностям, не требующую единого для всех модальностей векторного представления.

6. На основе разработанной меры расстояния построить модель кластеризации, позволяющую применять алгоритмы кластеризации и приближенного поиска ближайших соседей к мультимодальным данным без использования моделей глубокого обучения.

Методология и методы исследования. В качестве методологической базы исследования использовались существующие работы по алгоритмам приближенного поиска ближайших соседей, построению индекса в векторных базах данных и кластерному анализу (из областей машинного обучения и анализа данных), компьютерное моделирование, методы математической статистики (включая оценку распределений и тесты значимости), системного анализа, теории размещения и теории оптимизации (из сферы исследования операций и оптимизационных задач).

Новые научные результаты:

1. Предложен новый алгоритм классификации запросов по уровню сложности для приближенного поиска ближайших соседей с использованием IVF-индекса, позволяющий определять требуемую для достижения целевого показателя полноты область поиска, на основе числа результативных кластеров после начального этапа поиска.

2. Предложен новый адаптивный алгоритм поиска ближайших соседей в векторной базе данных на основе IVF-индекса, отличающийся от известных использованием классификатора сложности запросов на основе результатов предварительного поиска, использование которого позволяет повысить среднюю эффективность поиска.

3. Предложены новые эволюционные алгоритмы решения задачи k-средних, отличающиеся от известных оператором мутации, основанным на ускоренной жадной агломеративной процедуре, и позволяющие повысить точность решения задачи кластеризации.

4. Предложена новая модель кластеризации мультимодальных данных, которая, в отличие от известных моделей, позволяет напрямую, без приведения к единому векторному виду, применять алгоритмы кластеризации к мультимодальным данным.

Положения, выносимые на защиту:

1. Предложен классификатор сложности запросов, позволяющий оценить число кластеров, поиск в которых в среднем обеспечивает требуемое значение полноты (Recall) для каждого класса запросов. Классификатор обеспечивает точность (accuracy) определения класса сложности запроса на уровне 0,81.
2. Новый адаптивный алгоритм поиска ближайших на основе IVF-индекса обеспечивает ускорение выполнения запросов на 10–30% на наборах данных до 1 миллиарда объектов.
3. Новые алгоритмы решения задачи *k*-средних позволяют более точно решать задачу *k*-средних, за счет чего обеспечивается построение IVF-индекса, который позволяет ускорить выполнение запросов на 0,5-1%.
4. Новая модель автоматической группировки мультимодальных данных позволяет напрямую применять классические алгоритмы кластеризации с эффективностью не менее 70% (по индексу Рэнда) без построения дополнительных к существующим векторным представлениям модальностей структур данных.

Теоретическая значимость результатов. Теоретическая значимость работы состоит в развитии методов автоматической группировки объектов, развитии методов поиска данных в векторных базах данных, а также в расширении инструментария методов кластерного анализа для мультимодальных данных, в том числе больших мультимодальных данных.

Практическая значимость результатов. Предложенные модели и алгоритмы могут применяться для задач разделения объектов на группы и поиска данных в информационных системах, работающих с разнородными наборами данных в прикладных областях. Разработанный эволюционный алгоритм и адаптивный алгоритм поиска могут эффективно применяться в системах управления векторными базами данных. Использование эволюционных алгоритмов типа 1+λ, а также упрощенного жадного алгоритма позволяет снизить долю неверно определенных ближайших соседей в векторных базах данных, хранящих эмбеддинги. Разработанный адаптивный алгоритм приближенного поиска ближайших соседей и классификатор сложности запросов применяются в векторных базах данных с сотнями миллионов объектов для приближенного поиска ближайших соседей. Разработанная мера расстояния в пространстве мультимодальных данных позволяет вычислять расстояния между мультимодальными объектами, что позволяет не только адаптировать существующие алгоритмы кластеризации для такого пространства, но и сравнивать расстояния между собой, делая возможным приближенный поиск ближайших соседей в пространстве мультимодальных данных.

Исследование было выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания №FEFE-

2020-0013 «Развитие теории самоконфигурирующихся алгоритмов машинного обучения для моделирования и прогнозирования характеристик компонентов сложных систем», при поддержке Мегагранта «Гибридные методы моделирования и оптимизации в сложных системах»№075-15-2022-11-21, а также гранта Фонда Содействия Инновациям по программе «Код-ИИ» и хозяйственного договора с техкомпанией «Хуавей» (разработка эффективного алгоритма поиска объектов в векторной базе данных).

Апробация. Основные положения и результаты диссертационной работы докладывались на международных конференциях и семинарах: научный семинар Омского филиала института математики им. С.Л. Соболева СО РАН «Математическое моделирование и дискретная оптимизация» (2026), International Workshop on Mathematical Models and their Applications (IWMMA 2025), Hybrid methods of modeling and optimization in complex systems (HMMOCS 2022, 2024), 2021 3rd International Conference on Advanced Information Science and System, AISS 2021, 2020 International Conference on Control, Robotics and Intelligent System, CCRIS 2020.

Публикации. По теме диссертации опубликовано 24 работы, из них 6 публикаций входит в перечень ВАК РФ, 14 – в международных изданиях, индексируемых в системах цитирования Web of Science и Scopus. Зарегистрированы 2 программных системы в Федеральной службе по интеллектуальной собственности (Роспатент).

Структура работы. Диссертационная работа изложена на 142 страницах и состоит из введения, четырех глав, заключения, списка литературы из 150 источников, 23 таблиц, 26 рисунков и двух приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулирована цель работы и задачи, необходимые для ее достижения. Сформулированы научная новизна и практическая значимость работы, а также положения, выносимые на защиту.

В **первой главе** рассматривается задача поиска ближайших соседей (Nearest Neighbor Search) для векторных представлений (эмбеддингов): имеется фиксированное множество объектов (векторов) в векторной базе данных и множество векторов-запросов, для которых требуется быстро находить ближайшие объекты из базы данных по выбранной метрике. Точный поиск, то есть полный перебор, часто оказывается слишком затратным по времени, особенно в пространствах высокой размерности и на больших объемах данных, поэтому на практике широко применяются приближенные методы (Approximate Nearest Neighbor, ANN), допускающие ошибки, однако работающие на несколько порядков быстрее.

Далее вводится ключевая идея построения индекса для приближенного поиска ближайших соседей: заранее выполнить предварительную обработку (индексацию) набора векторов так, чтобы во время запросов получать результат быстрее точного поиска. В качестве контекста упоминаются распространенные направления ANN-поиска и индексации: деревья (например, варианты k-d деревьев), хеширование

(локально-чувствительное хеширование) и подходы, основанные на квантизации векторов, то есть упрощенном их представлении. Отдельно отмечается практическая значимость для задач поиска по похожести в системах обработки разнородных (в том числе мультимодальных) данных, где объекты представлены унифицированными векторами.

Приводится описание подхода, основанного на IVF-индексе, где индекс строится как набор кластеров, а поиск сводится к двум стадиям:

1. Выбрать небольшое число наиболее перспективных кластеров для запроса.
2. Выполнить полный перебор векторов, потенциально являющихся ближайшими соседями, только внутри выбранных кластеров.

Перспективность кластера определяется как расстояние до его характерного вектора, то есть центра или центроида, по заданной мере расстояния (например, евклидовой или косинусной).

Вторая глава посвящена адаптивному алгоритму приближенного поиска ближайших соседей в векторных базах данных. Разработанный алгоритм позволяет находить ближайших соседей с высокой точностью (значение Recall до 0.99). Вводится понятие «сложности» запроса. Под сложностью запроса понимается количество кластеров, которое необходимо обработать (*nprobe*) для достижения требуемой точности. Выдвигается гипотеза о том, что сложность запроса можно определить на ранних стадиях поиска, то есть после обработки некоторого минимального числа кластеров. После выполнения поиска в этом минимальном числе кластеров можно определить некоторые признаки запроса, наиболее результативным из которых является *nres* – число результативных кластеров, то есть кластеров, в которых нашелся хотя бы один ближайший сосед. Отмечается, что дистанционные статистики, такие как средние расстояния до соседей или центроидов, не являются полезными и информативными, так как в пространстве большой размерности расстояния слабо отличаются между собой.

Итоговый алгоритм адаптивного поиска ориентируется только на параметр *nres* и опирается на заранее обученный классификатор сложности запроса.

Для обучения классификатора требуется набор размеченных данных: каждому вектору-запросу из обучающей выборки ставится в соответствие набор реальных ближайших соседей. Поиск реальных ближайших соседей требует больших вычислительных затрат, однако процесс обучения классификатора не выполняется для каждого нового запроса. Обучить классификатор необходимо только один раз (при построении индекса) на небольшой обучающей выборке из нескольких сотен запросов. Классификатор определяет сложность запроса по единственному параметру: *nres*. Несмотря на простоту, разработанный классификатор определяет класс сложности запроса к векторной базе данных с точностью до 0,81, при этом практически не требуя вычислительных затрат для определения расстояний между объектами или любых других параметров, помимо числа результативных кластеров. Обученный классификатор по полученному после обработки минимального числа

кластеров значению *nres* определяет общее число кластеров, в которых нужно произвести поиск ближайших соседей, чтобы найти ближайших соседей с заданной точностью. Класс сложности определяется набором значений *nres*.

В работе используются четыре класса сложности. Первый класс сложности – это запросы, для которых достаточно обработать лишь некоторое заданное минимальное число кластеров, чтобы определить ближайших соседей с достаточной точностью, то есть для таких запросов не требуется продолжать поиск. Остальные три класса определяются путем разбиения запросов в обучающей выборке, которые не вошли в первый класс сложности, на равномощные подмножества по значению *nres*. Минимальное и максимальное значение *nres* в каждом классе определяет границы этого класса. Алгоритм обучения классификатора представлен в алгоритме 1, алгоритм адаптивного приближенного поиска ближайшего соседа представлен в алгоритме 2.

Алгоритм 1 - Алгоритм обучения классификатора сложности запросов

Дано: Обучающая выборка запросов Q, представленная в виде случайной выборки векторов данных; ожидаемое значение полноты Recall@K; количество ближайших соседей K; минимальное количество проверяемых кластеров $n_{minchecked}$.

Шаг 1. Для каждого запроса q из набора Q выполнить исчерпывающий поиск для определения истинных ближайших соседей;

Шаг 2. Для каждого запроса q_i в наборе Q выполнить поиск в $n_{minchecked}$ кластерах и рассчитать количество найденных результатов n_{res};

Шаг 3. Для каждого запроса q_i в наборе Q выполнять:

Шаг 4. Установить начальное количество проверяемых кластеров $nprobe=0$;

Шаг 5. Увеличить значение *nprobe* на единицу ($nprobe \leftarrow nprobe+1$);

Шаг 6. Оценить значение полноты Recall(q_i)@K, полученное после сканирования *nprobe* кластеров;

Шаг 7. Если Recall(q_i)@K ≥ Recall@K, то вернуться к шагу 5;

Шаг 8. Зафиксировать необходимое количество кластеров для данного запроса: $nprobe@q_i=nprobe$;

Шаг 9. Завершить цикл для текущего запроса;

Шаг 10. Разделить набор Q на четыре подмножества на основе $nprobe@q_i$ и определить границы классов сложности следующим образом:

Шаг 10.1. Установить $M1=n_{minchecked}$;

Шаг 10.2. Сформировать множество *Class*1, включающее запросы, где $nprobe@q_i \leq n_{minchecked}$;

Шаг 10.3. Определить остаточное множество $Classes234=Q \setminus Class1$;

Шаг 10.4. Рассчитать $M2$ как 33-й процентиль значений *nprobe* среди запросов из *Classes*234;

Шаг 10.5. Рассчитать $M3$ как 66-й процентиль тех же значений;

Шаг 10.6. Сформировать множества классов сложности: $Class2 \leftarrow M1 < nprobe@q_i \leq M2$; Class3 $\leftarrow M2 < nprobe@q_i \leq$ M3; Class4 $\leftarrow nprobe@q_i >$ M4;

Шаг 11. Для каждого класса сложности $i \in \{1,2,3,4\}$ рассчитать среднее значение $nprobe@q_i$, при котором достигается требуемая полнота, и сохранить их как параметры $S1,S2,S3,S4$.

Шаг 12. Вернуть полученные значения порогов классов и значения *nprobe* для каждого класса: $M1,M2,M3,S1,S2,S3,S4$.

Результатом работы алгоритма 1 являются значения $M1$, $M2$, $M3$, $S1$, $S2$, $S3$, $S4$.

Алгоритм 2 - Адаптивный алгоритм поиска приближенных ближайших соседей

Дано: Вектор запроса q, количество искомых ближайших соседей K, минимальное количество проверяемых кластеров $n_{minchecked}$ границы классов сложности $M1,M2,M3$, значения количества проверяемых кластеров *nprobes* для каждого класса сложности $S1,S2,S3,S4$, обученный классификатор для K, $n_{minchecked}$ и ожидаемой Recall@K (см. алгоритм 2.1).

Шаг 1. Проверить наличие дополнения к индексу (обученного классификатора) для заданных K, $n_{minchecked}$ и ожидаемой Recall@K. Если классификатор не обучен: выполнить стандартную процедуру поиска ближайших соседей (ANN) и завершить алгоритм;

Шаг 2. Выполнить поиск K ближайших соседей в фиксированном минимальном количестве кластеров $n_{minchecked}$ алгоритмом IVF-поиска;

Шаг 3. Рассчитать количество найденных результатов n_{res};

Шаг 4. Классифицировать сложность запроса по количеству результатов n_{res} и выбрать соответствующее значение *nprobes*:

Если $n_{res} \leq M1$, то $nprobes \leftarrow S1$;

Если $M1 < n_{res} \leq M2$, то $nprobes \leftarrow S2$;

Если $M2 < n_{res} \leq M3$, то $nprobes \leftarrow S3$;

Если $n_{res} > M3$, то $nprobes \leftarrow S4$;

Шаг 5. Выполнить дополнительный поиск алгоритмом IVF-поиска в ($nprobes$ - $n_{minchecked}$) дополнительных кластерах;

Шаг 6. Вернуть найденных K ближайших соседей.

Процесс поиска ближайших соседей начинается с получения вектора запроса q, указанного количества ближайших соседей K и ожидаемого уровня полноты. Затем, алгоритм проверяет, обучен ли классификатор для этих значений. Если классификатор не обучен, то запускается стандартная процедура приближенного поиска ближайших соседей. В том случае, если классификатор обучен, алгоритм ищет K ближайших соседей в заданном минимальном количестве кластеров. Окончательное число кластеров, в которых нужно провести поиск *nprobe*, уточняется по результатам этого поиска.

Несмотря на кажущуюся простоту реализации, такой алгоритм показал значительный прирост производительности (таблица 1). В таблице 1 Latency – это время отклика, QPS – количество обработанных запросов в секунду. Эксперименты проведены на устройстве с центральным процессором 4 x Kunpeng-920 5250, устройством NPU 8 x Ascend 910B42NPU, ОЗУ 1.5ТБ DDR4, SSD-диск 7ТБ NVMI. В каждом эксперименте запросы распределялись между 100 параллельными процессами.

Таблица 1 - Производительность приближенного поиска ближайших соседей

Количество векторов данных	Показатель	Без адаптивного классификатора	С адаптивным классификатором	Прирост, %
10^7	Recall, %	99,12	99,12	0%
	QPS, c^{-1}	335,61	422,13	+25,78%
	Latency	296,53	235,17	-20,68%
10^8	Recall, %	98,63	99,01	+0,39%
	QPS, c^{-1}	136,21	145,21	+6,6%
	Latency	719,8	670,82	-6,8%
10^9	Recall, %	99,04	99,05	+0,1%
	QPS, c^{-1}	6,08	8,47	+39.31%
	Latency	7959	5607	-29.55%

Значительный прирост скорости обработки запросов позволяет сделать вывод об эффективности разработанного адаптивного алгоритма поиска приближенных ближайших соседей. Также достоинством алгоритма является простота его реализации и внедрения в уже существующие программные решения.

Третья глава посвящена эволюционным алгоритмам для создания IVF-индекса с оператором мутации, основанным на ускоренной жадной агломеративной процедуре. Алгоритм решает задачу *k*-средних или *p*-медиан, которая также может рассматриваться и как задача автоматической группировки данных (кластеризации), так и как задача размещения. Задача состоит в том, чтобы расположить *p* векторов таким образом, чтобы сумма расстояний в заданной мере (функции) расстояния от каждого объекта до ближайшего вектора из числа искомых *p* векторов была минимальной. Задача *p*-медиан, в которой расстояния от объектов до центров вычисляются как квадрат Евклидова расстояния, идентична задаче *k*-средних. Для задач создания IVF-индекса *k* найденных векторов являются центрами кластеров, аппроксимирующими объекты, принадлежащие этому кластеру, а область поиска ближайших соседей определяется близостью вектора-запроса к центрам кластеров. Классическим алгоритмом решения задачи кластеризации является процедура Ллойда (алгоритм 3).

Алгоритм 3 - Процедура Ллойда

Дано: векторы данных $A_1...A_N$, k начальных центров кластеров $X_1...X_k$.

Шаг 1. Составить кластер C_i векторов данных для каждого центра X_i так, чтобы для каждого вектора данных его центр был ближайшим;

Шаг 2. Рассчитать новое значение центра X_i для каждого кластера;

Шаг 3. Если Шаги 1-2 не привели к изменениям, то ОСТАНОВ, иначе переход к Шагу 1.

Эволюционный алгоритм (1+λ) генерирует λ новых решений на каждой итерации с использованием оператора мутации, где лучшее из них может на каждой итерации заменить исходное. Параметр λ определяется как количество создаваемых потомков на каждой итерации. В настоящей работе предлагается новый эволюционный алгоритм (1+λ) для задач кластеризации и размещения, использующий специальный оператор мутации на основе кроссовера с жадной агломеративной процедурой (алгоритм 4).

Алгоритм 4 - $(1+\lambda)$ эволюционный алгоритм с жадной агломеративной кроссовероподобной мутацией

Дано: Набор объектов A, число центров k, количество создаваемых потомков на каждой итерации λ.

Шаг 1. Случайным образом создать подмножество $S \subset \{A_1,\ldots,A_N\}$, $A_i \in A$; $r \leftarrow [p/2]$; $S \leftarrow ALA(S)$.

Шаг 2. Если $r<p/4$, то генерируется λ решений потомков и переход к шагу 3, иначе генерируется только одно решение и переход к шагу 12.

Шаг 3. Для каждого i из {1,...,λ} выбрать случайное подмножество$i' \subset \{A_1, \ldots, A_N\}$; $S_{i'} \leftarrow ALA(S_{i'})$, если $i \leq [\lambda/2]$, то r' выбрать из$\left\{\overline{\max\left(1, \frac{r-1}{2}\right), r}\right\}$, иначе из$\{\overline{r+1, [2(r+1)]}\}$; $S_i' \leftarrow AGGLr'(S, S_{i'})$.

Шаг 4. Определить i, обеспечивающее минимальное значение $F(S_i')$.

Шаг 5. Если $F(S_i') < F(S)$ и $i \leq [\lambda/2]$, то $r \leftarrow \max\left\{1, \frac{r}{1.5}\right\}$.

Шаг 6. Если $F(S_i') < F(S)$и $i > [\lambda/2]$, то $r \leftarrow \min\left\{\left[\frac{k}{2}\right], 1.5r\right\}$.

Шаг 7. Если $F(S_i') > F(S)$ и $i \leq [\lambda/2]$, то $r \leftarrow \max\left\{1, \frac{r}{1.25}\right\}$.

Шаг 8. Если $F(S_i') > F(S)$ и $i > [\lambda/2]$, то $r \leftarrow \max\{1, 1.25r\}$.

Шаг 9. Если $r=1$, то с вероятностью 0.1 выполнить $r \leftarrow \left[\frac{p}{2}\right]$.

Шаг 10. Если $r=[p/2]$, то с вероятностью 0.1 выполнить $r \leftarrow 1$.

Шаг 11. Перейти к Шагу 15.

Шаг 12. Случайным образом определить $S \subset \{A_1,\ldots,A_N\}$; $S' \leftarrow ALA(S')$.

Шаг 13. Случайным образом определить r' из $\left\{\overline{\max\{1, \left[\frac{r-1}{1.5}\right], [1.5(r+1)]\}}\right\}$; $S \leftarrow AGGLr'(S, S')$.

Шаг 14. Если $F(S') < F(S)$, то $S \leftarrow S'$.

Шаг 15. Если не исчерпано доступное на выполнение время, то вернуться к Шагу 2.

Шаг 16. ОСТАНОВ.

Здесь ALA – запуск процедуры Ллойда, AGGL – запуск жадной агломеративной процедуры, которая, начиная с решения, содержащего избыточное число центроидов, исключает часть центров (центроидов) из решения, доводя их количество до требуемого значения *k*.

На рисунке 1 проиллюстрированы результаты экспериментов, проведенных на наборе данных BIRCH.

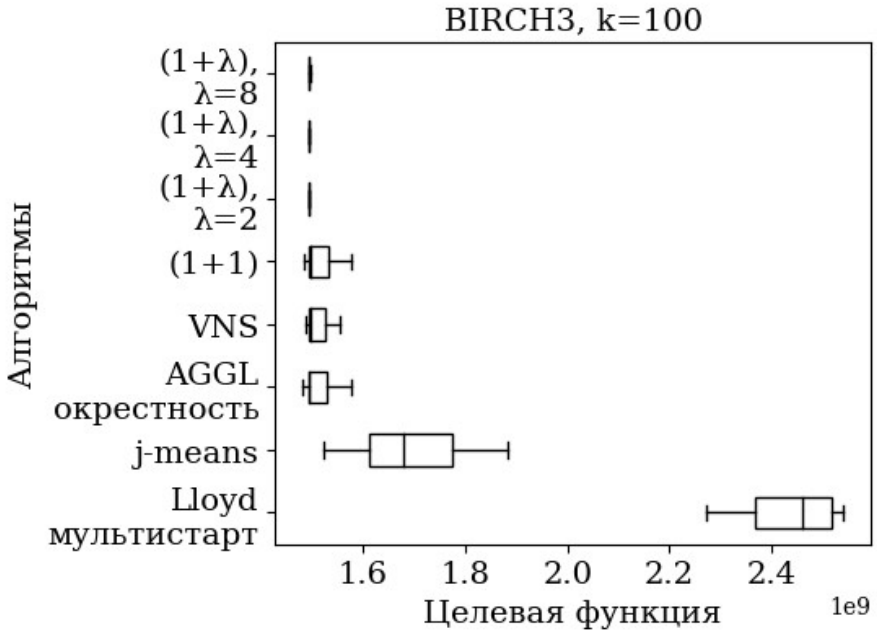

Рисунок 1 - Сравнение результатов работы (1+λ) эволюционного алгоритма с кросс-мутацией с другими алгоритмами кластеризации, коробчатая диаграмма. Границы диаграммы указывают на наблюдаемый минимум (слева) и наблюдаемый максимум (справа), границы коробки верхний и нижний квартиль, а линия внутри коробки обозначает медиану

Результаты экспериментов подтверждают эффективность разработанного (1+λ) эволюционного алгоритма, который позволяет получать стабильный и хороший результат по значению целевой функции. Статистическая значимость результатов подтвержается *t*-критерием и критерием Манна-Уитни-Уилкоксона.

В многомерных пространствах данных (особенно при работе с эмбеддингами) как внутрикластерные, так и межкластерные расстояния, как правило, распределены в

узком диапазоне (все расстояния, как внутрикластерные, так и межкластерные, почти одинаковы). Это явление является следствием «проклятия размерности» и приводит к выравниванию распределений расстояний. Предполагается, что увеличение целевой функции алгоритма Ллойда (суммы расстояний или квадратов расстояний) при удалении центроида и перераспределении объектов его кластера между другими центроидами зависит только от мощности (размера) удаляемого кластера. На основе этого предположения разработана специализированная агломеративная процедура, направленная на автоматическую группировку объектов в многомерных пространствах. Эта процедура значительно ускоряет построение индексов инвертированного файла (IVF) по сравнению с классическими агломеративными подходами: вместо прямого вычисления приращения целевой функции предложенный метод аппроксимирует его, используя мощность удаляемого кластера (см. алгоритм 5). Эта аппроксимация позволяет применять агломеративные подходы к наборам данных высокой размерности и большого объема, для которых такие методы обычно нецелесообразны из-за их высокой вычислительной сложности.

Алгоритм 5 - Ускоренная агломеративная процедура (AGGL) для кластеризации эмбеддингов

Дано: Исходное решение задачи кластеризации S с начальным числом центроидов k_0, требуемое число кластеров и центроидов k, $k_0>k$, число кластеров, удаляемое за одну итерацию k_e.

Шаг 1. Выполнить алгоритм Ллойда с k_0 кластерами: $S \leftarrow ALA(S)$.

Шаг 2. $k' \leftarrow k_0$.

Шаг3. $k' \leftarrow \begin{cases} k' - k_e, & k - k_e \geq k \\ k, & k - k_e < k' \end{cases}$.

Шаг 4. Удалить k' центроидов, соответствующих кластерам наименьшей мощности, перераспределить векторы данных по центроидам (переформировать кластеры).

Шаг5.Повторять Шаги 3-4 пока $k' \neq k$.

Предлагаемая процедура, несмотря на свою простоту, повышает качество формируемого IVF-индекса. В результате улучшаются как целевые показатели модели кластеризации (например, сумма квадратов расстояний до центроидов для k-средних или сумма расстояний до центров для p-медиан), так и эффективность поиска в векторной базе данных при использовании индекса, построенного на основе новых алгоритмов.

Для оценки производительности поиска в векторной БД с использованием IVF-индекса был проведен ряд экспериментов в контролируемых условиях. Индекс строился как с применением стандартного алгоритма k-средних, так и с использованием новых алгоритмов. В этих экспериментах были предприняты усилия для достижения сопоставимых значений Recall@K в различных конфигурациях за счет регулирования числа обрабатываемых в ходе выполнения запроса к векторной

БД с построенным IFV-индексом кластеров (т.е. за счет регулирования значения *nprobes*), что позволило провести содержательное сравнение показателя Queries per Second (QPS – число обработанных запросов в секунду), который служит ключевым индикатором эффективности и пропускной способности системы в различных условиях. Параметр *nprobes* определяет количество ближайших центроидов для поиска ближайших соседей.

Таблица 2 - Сравнительные результаты поиска по IVF-индексам на основе различных алгоритмов кластеризации

	Алгоритм Ллойда, набор данных (10^5 128-мерных векторов) SIFT100K, 1024 кластера, *nprobes*= 99	Новые ускоренные алгоритмы, набор данных SIFT100K (10^5 128-мерных векторов), 1024 кластера, *nprobes* = 98	Новые ускоренные алгоритмы, SIFT10M набор данных (10^7 128-мерных векторов), 3150 кластеров, *nprobes* = 125
Recall@100, среднее	99.08	**99.09**	99.21
Recall@100, медианное	99.09	99.09	99.21
Recall@100, станд.откл.	0.0196	**0.0042**	0.0095
QPS, среднее	640.11	**658.95**	307.84
QPS, медианное	641.59	**661.99**	305.09
QPS, станд.откл.	27.12	**10.44**	38.22
Задержка, среднее, мкс	155.03	**151.03**	327.36
Задержка, медианная, мкс	155.12	**150.32**	327.2
Задержка, станд. откл., мкс	5.40	**2.46**	40.55

Как показывают данные в таблице 2, новая жадная агломеративная процедура обеспечивает построение индекса IVF, повышающего точность поиска. В частности, при поиске в одинаковом количестве кластеров (*nprobe*) этот подход дает более высокое среднее значение Recall, что указывает на большую долю релевантных результатов по сравнению с традиционным методом (алгоритм Ллойда). Это улучшение обусловлено способностью процедуры более эффективно оптимизировать назначение кластеров, уменьшая ошибки в приближениях ближайших соседей и

обеспечивая более точные совпадения запросов без чрезмерных вычислительных ресурсов.

Кроме того, процедура не только повышает точность, но и ускоряет общий процесс поиска, что подтверждается увеличением количества запросов в секунду (QPS) и уменьшением показателей задержки. Помимо этих преимуществ в скорости и точности, метод обеспечивает значительно большую стабильность результатов по производительности, проявляющуюся в стабильном качестве поиска и времени выполнения на различных наборах данных и при различной нагрузке запросов. Такая надежность особенно ценна в реальных условиях, где колебания производительности могут подорвать доверие пользователей и эффективность системы, что делает жадный агломеративный подход многообещающим шагом вперед в масштабируемых методах индексирования.

В **четвертой главе** представлена новая модель кластеризации мультимодальных данных, в основе которой лежит особый способ агрегации модальностей для вычисления расстояний (рисунок 2) между объектами, что позволяет применять известные алгоритмы кластеризации, использующие произвольные меры расстояния, к мультимодальным данным и строить индексы векторных баз данных на основе кластеризации. Под мультимодальными данными понимаются данные, представленные разнородными видами информации: каждый объект может одновременно описываться вектором, текстом и изображением.

Рисунок 2 – Основные компоненты модели кластеризации мультимодальных данных

Перед расчетом расстояний необходимо привести все представления (изображения, тексты и т.д.) в векторный вид.

Предполагается, что весовые коэффициенты модальностей определяются экспертно (например, равными) или согласно определенной экспертом эвристике.

Для нормализации можно использовать различные подходы, включая нормализацию по среднеквадратическому отклонению, более устойчивую к аномалиям в данных (выбросам), но для предобработанных данных и эмбеддингов стандартной практикой является нормализация в диапазон от 0 до 1 (Min-Max Scaling).

Для агрегации модальностей использовалась формула

$$D(X_i, X_j) = \sqrt{\sum_{k=1}^{m} \alpha_k \ \hat{d}_k^{\,2}\left(X_i^k, X_j^k\right)},$$

где $D(X_i, X_j)$ - расстояние между мультимодальными объектами, α_k - весовой коэффициент k-й модальности, $\hat{d}_k$ - нормированное внутримодальное расстояние. Существует возможность использования других способов агрегации модальностей. Метод конкатенации векторов приводит к тому, что модальности высокой размерности слишком сильно влияют на конечное расстояние между объектами. Метод ADAPT позволяет обрабатывать мультимодальные данные, не прибегая к глубокому обучению, но он также имеет ряд недостатков: в современных задачах зачастую требуется обработка эмбеддингов, то есть векторов действительных чисел высокой размерности. Метод ADAPT приводит все действительные числа к конечному множеству дискретных значений (фактически осуществляя квантизацию), не уменьшая размерность данных, но приводя к потере информации, что приводит к потере точности без повышения вычислительной эффективности.

Предлагаемая модель позволяет проводить кластеризацию данных, которые описываются разнородными данными, без использования размеченных данных. Это позволяет модифицировать некоторые алгоритмы кластеризации для решения задачи автоматической группировки мультимодальных данных.

В отличие от подходов, основанных на глубоком обучении, разработанная модель не требует никакой, даже минимальной разметки данных, не требует обучения и дообучения. Для составления эмбеддингов могут применяться любые модели, как широко известные и распространенные, такие как BERT, так и произвольные, разработанные пользователем.

Были реализованы следующие алгоритмы кластеризации: k-means, Meanshift, DBSCAN, спектральный алгоритм с матрицей схожести, спектральный алгоритм с матрицей смежности, BIRCH, Bisecting K-Means. Алгоритмы кластеризации с новой мерой расстояния были реализованы на Apache Spark и могут быть использованы в распределенных вычислительных системах.

Был проведен ряд вычислительных экспериментов на различных наборах мультимодальных данных размером до 10Гб, с числом модальностей до пяти, каждая модальность описана вектором размерности не больше 128. Для таких наборов данных среднее достигнутое значение индекса Рэнда составило 0,78.

Рисунок 3 иллюстрирует зависимость достигнутого качества кластеризации (по мере SF) в зависимости от времени, затраченном на оптимизацию гиперпараметров алгоритмов кластеризации. Использовался алгоритм автоматического выбора и настройки алгоритмов кластеризации MASSCAH. Решалась задача кластеризации

археологических находок. SF (Score Function) – это мера качества кластеризации, в которой разделимость измеряется на основе расстояния от центроидов кластеров до глобального центроида, а компактность – на основе расстояния от точек внутри кластера до его центроида. Мера SF вычисляется следующим образом:

$$SF(C) = \frac{\sum_{c_k \in C} \min\limits_{c_l \in C \setminus c_k} \{\|\bar{c_k} - \bar{c_l}\|\}}{\sum_{c_k \in C} 10/|c_k| \sum \max\limits_{x_i \in c_k}(0.1 * |c_k|) * \|\bar{x_i} - \bar{c_k}\|},$$

где C – полученное разбиение объектов на кластеры. Каждый объект в наборе данных описывает исторический артефакт (нательный крест), найденный археологами в определенном месте. Первая модальность описывает место находки (координаты), вторая описывает химический состав металла (количество различных примесей в серебре), третья модальность – векторное представление фотографии найденного исторического артефакта. Для каждой модальности использовалось евклидово расстояние, всего 100 объектов.

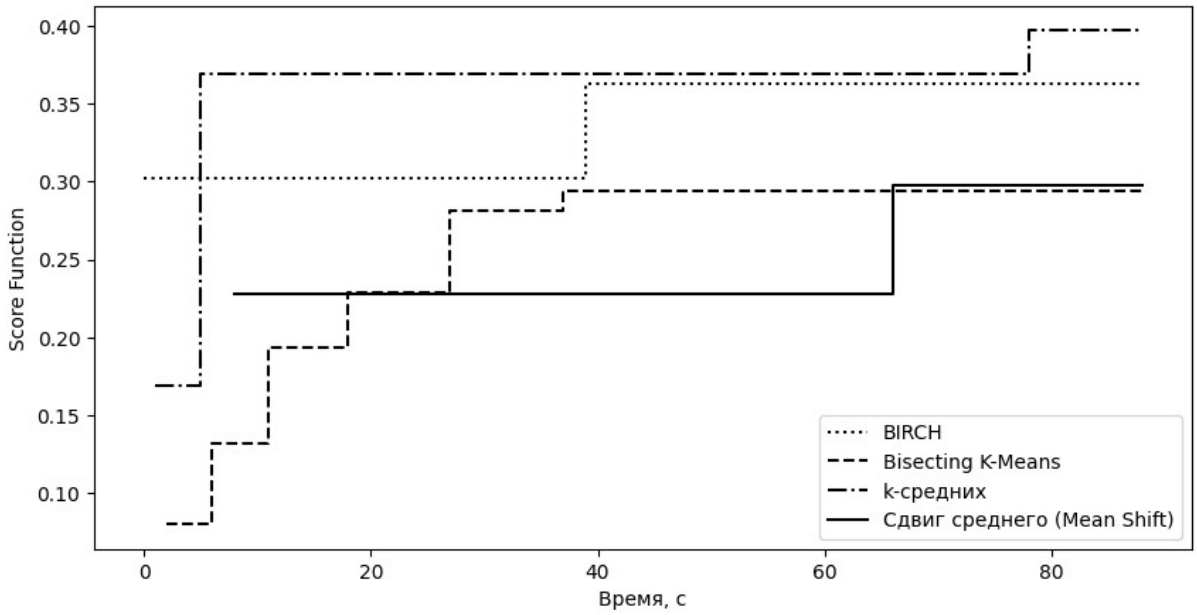

Рисунок 3 - Результат кластеризации археологических находок, представленных как мультимодальные данные. По вертикали - достигнутая SF мера, по горизонтали - время, с. Каждая линия на графике – результат работы алгоритма оптимизации гиперпараметров алгоритмов кластеризации Optuna для отдельного алгоритма кластеризации

Разработанные на основе новой меры расстояния между мультимодальными объектами алгоритмы кластеризации применимы к самому широкому кругу задач с различными модальностями разной размерности. Для подтверждения применимости

разработанной меры расстояния был проведен ряд экспериментов на более чем десяти наборах размеченных данных, как реальных, так и сгенерированных. Для выбора алгоритма кластеризации и настройки его гиперпараметров использовался алгоритм MASSCAH. Средние достигнутые внешние меры качества кластеризации в сравнении с методом ADAPT представлены в таблице 3. Результаты таблицы показывают, что разработанная модель кластеризации способна определять группы схожих объектов на неразмеченных данных лучше.

Таблица 3 -Средние значения достигнутых внешних мер качества кластеризации для нескольких наборов размеченных данных.

Мера качества	Агрегация модальностей, среднее значение меры качества на сгенерированных данных	Агрегация модальностей, среднее значение меры качества на реальных данных	ADAPT, среднее значение меры качества на сгенерированных данных	ADAPT, среднее значение меры качества на реальных данных
Rand Index, среднее	0.78	0.73	0.68	0.66
Jaccard Index, среднее	0.69	0.71	0.63	0.63
F-мера, среднее	0.72	0.75	0.65	0.66

Разработанная модель кластеризации мультимодальных данных позволяет эффективно разделять объекты на группы, а также применять алгоритмы автоматической группировки данных, на что указывает значение SF меры на рисунке 3 и значения внешних мер качества кластеризации в таблице 3. Преимущества разработанной меры расстояния заключаются в вычислительной простоте, отсутствии необходимости применять модель глубокого обучения для агрегации модальностей и создании единого векторного пространства, а также в широких возможностях для модификаций. Алгоритмы кластеризации с разработанной мерой расстояния могут использоваться в распределенных вычислительных системах.

Заключение

В диссертационной работе предложены алгоритмы автоматической группировки данных для построения IVF-индекса и адаптивный алгоритм поиска ближайших соседей, позволяющие повысить точность и скорость работы алгоритмов поиска ближайших соседей на основе IVF-индекса на 10-39%. Были решены следующие задачи:

1. Представлен алгоритм классификации сложности запросов приближенного поиска ближайших соседей с использованием IVF-индекса, позволяющий определить сложность запроса и предсказать количество кластеров, в которых находятся ближайшие соседи, а также алгоритм для обучения этого классификатора. Продемонстрировано значительное повышение производительности поиска. Продемонстрирована применимость такого алгоритма к большим объемам данных.
2. Разработан адаптивный алгоритм поиска в векторной базе данных на основе IVF-индекса, который по результатам предварительного поиска и классификации сложности запроса динамически определяет ограничения пространства поиска и повышает вычислительную эффективность при сохранении требуемой точности, прирост составил 10-39% на наборах данных объемом до 10^9 объектов при значении полноты 0,99.
3. Разработаны новые эволюционные алгоритмы решения задачи k-средних, отличающиеся от известных применением оператора мутации, основанном на ускоренной жадной агломеративной процедуре. Результаты вычислительных экспериментов подтверждают стабильность и качество кластеризации, достигаемое новым эволюционным алгоритмом, а также прирост вычислительной эффективности процедуры поиска приближенных ближайших соседей по IVF-индексу.
4. Представлена новая модель кластеризации мультимодальных данных, позволяющая применять алгоритмы кластеризации напрямую, без приведения модальностей к единому векторному пространству и повысить точность решения задачи кластеризации.

В настоящем диссертационном исследовании была разработана методология ускорения поиска в векторных базах данных, включающая определение глубины поиска, оптимизацию построения индекса методами кластеризации и обработку сложных типов данных (мультимодальных данных). Предложенные алгоритмы обеспечивают практическое повышение производительности при сохранении заданного качества поиска и демонстрируют устойчивость на больших объемах данных. Сформулированные подходы могут быть использованы при проектировании и модернизации высокопроизводительных систем приближенного поиска.

Также представленные в настоящей работе исследования открывают несколько направлений для дальнейших исследований. Оценка доли результативных кластеров на раннем этапе поиска ближайших соседей для определения глубины поиска может быть развита в дальнейшем, как для адаптивных методов поиска, так и для других задач. Принцип аппроксимации прироста целевой функции задачи k-средних, использованный в ускоренной агломеративной процедуре кластеризации

эмбеддингов, может найти применение и в других практических задачах, включающих кластеризацию. Новая модель кластеризации мультимодальных данных, не использующая моделей глубокого обучения, также демонстрирует многообещающие перспективы практического применения.

Список опубликованных работ по теме диссертации.

Основные результаты диссертационной работы опубликованы в научных журналах и материалах конференций. Всего опубликовано 24 работы, из которых 6 входит в перечень ВАК.

Публикации, входящие в журналах «Белого списка»

1. Казаковцев, В. Л. Комбинация жадной агломеративной эвристики и эволюционного алгоритма для задачи размещения / В. Л. Казаковцев // Системы управления и информационные технологии. – 2024. – № 1(95). – С. 40-44. (ВАК К2 по специальности 2.3.1)

2. Казаковцев, В. Л. Об операторе мутации в эволюционном алгоритме автоматической группировки / В. Л. Казаковцев // Системы управления и информационные технологии. – 2022. – № 2(88). – С. 96-100. – DOI 10.36622/VSTU.2022.88.2.019. (ВАК К2 по специальности 2.3.1)

3. О нормализации данных в задаче автоматической группировки промышленной продукции по однородным производственным партиям / Ф. Г. Ахматшин, И. Р. Насыров, В. Л. Казаковцев, Л. А. Казаковцев // Системы управления и информационные технологии. – 2020. – № 2(80). – С. 86-89. (ВАК К2 по специальности 2.3.1)

4. Рожнов, И. П. Реализация жадных эвристических алгоритмов кластеризации для массивно-параллельных систем / И. П. Рожнов, В. Л. Казаковцев // Системы управления и информационные технологии. – 2019. – № 2(76). – С. 36-40. (ВАК К2)

5. Алгоритм для задачи к-средних с рандомизированными чередующимися окрестностями / И. П. Рожнов, Л. А. Казаковцев, М. Н. Гудыма, В. Л. Казаковцев // Системы управления и информационные технологии. – 2018. – № 3(73). – С. 46-51. (ВАК К2 по специальности 2.3.1)

6. Составление оптимальных ансамблей алгоритмов кластеризации / И. П. Рожнов, В. И. Орлов, М. Н. Гудыма, В. Л. Казаковцев // Системы управления и информационные технологии. – 2018. – № 2(72). – С. 31-35. (ВАК К2 по специальности 2.3.1)

7. Fast Adaptive Approximate Nearest Neighbor Search with Cluster-Shaped Indices / V. Kazakovtsev, M. Plekhanov, A. Naumchev [et al.] // Big Data and Cognitive Computing. – 2025. – Vol. 9, No. 10. – P. 254. – DOI 10.3390/bdcc9100254. (квартиль в WoS: 1, квартиль в Scopus: 1, уровень 1 «Белого списка»)

8. Algorithms with greedy heuristic procedures for mixture probability distribution separation / L. Kazakovtsev, D. Stashkov, M. Gudyma, V. Kazakovtsev // Yugoslav Journal

of Operations Research. – 2019. – Vol. 29, No. 1. – P. 51-67. – DOI 10.2298/YJOR171107030K. (квартиль в Scopus: 3, уровень 2 «Белого списка»)

9. Реализация алгоритма составления расписания детского центра / Е. Б. Пацук, Л. А. Казаковцев, И. Р. Насыров [и др.] // Современные наукоемкие технологии. – 2018. – № 8. – С. 132-137 (уровень 2 «Белого списка»).

Другие публикации, индексируемые в Web of Science и Scopus:

10.Kazakovtsev, V. Data segmentation through two-level clustering with greedy approach / V. Kazakovtsev, E. Markushin // ITM Web of Conferences : III International Workshop, Krasnoyarsk, 02–04 декабря 2024 года. – Krasnoyarsk: EDP Sciences, 2025. – P. 4007. – DOI 10.1051/itmconf/20257204007.

11.An opensource library for AutoML multimodal clustering on Apache Spark / S. B. Muravyov, V. L. Kazakovtsev, I. S. Usov [et al.] // Записки научных семинаров Санкт-Петербургского отделения математического института им. В.А. Стеклова РАН. – 2024. – Vol. 540. – P. 178-193.

12.Golovanov S.M. Determination of the Homogeneity of a Set of Elements on the Basis of the Quality Characteristics of the Division of a Set into Groups / S.M. Golovanov, V.L. Kazakovtsev, G.Sh. Shkaberina // The Tenth International Workshop on Mathematical Models and their Applications, Krasnoyarsk, the Russian Federation, November 16-18, 2021. – 2023.

13.Kazakovtsev, L. A (1 + λ) evolutionary algorithm with the greedy agglomerative mutation for p-median problems / L. Kazakovtsev, I. Rozhnov, V. Kazakovtsev // AIP Conference Proceedings : Proceedings of the IV International Scientific Conference on Advanced Technologies in Aerospace, Mechanical and Automation Engineering: (MIST: Aerospace-IV 2021), Krasnoyarsk, 10–11 December 2021 года. Vol. 2700. – AIP Publishing: AIP Pablishing, 2023. – P. 040003. – DOI 10.1063/5.0124952.

14.Gradient Neural Dynamics Based on Modified Error Function / P. S. Stanimirovic, D. Gerontitis, N. Tesic, V. Kazakovtsev [et al.] // Hybrid methods of modeling and optimization in complex systems : Proceedings of the International Workshop “Hybrid methods of modeling and optimization in complex systems” (HMMOCS’2022), Krasnoyarsk, 22–24 ноября 2022 года. – London, United Kingdom: European Proceedings, 2023. – P. 256-263. – DOI 10.15405/epct.23021.31.

15.Kazakovtsev, L. A. System for Automatic Grouping of Metadata of Three-Dimensional Models / L. A. Kazakovtsev, V. V. Kutsevalova, V. L. Kazakovtsev // Hybrid methods of modeling and optimization in complex systems : Proceedings of the International Workshop “Hybrid methods of modeling and optimization in complex systems” (HMMOCS’2022), Krasnoyarsk, 22–24 November 2022. – London, United Kingdom: European Proceedings, 2023. – P. 343-350. – DOI 10.15405/epct.23021.42.

16.Kazakovtsev, V. Application of the automatic selection and configuration of clustering algorithms method for the Apache Spark framework / V. Kazakovtsev, S. Muravyov // ACM International Conference Proceeding Series: 3, Sanya, 26–28 ноября 2021 года. – Sanya, 2021. – DOI 10.1145/3503047.3503104.

17.Recommender system for an academic supervisor with a matrix normalization approach / V. Kazakovtsev, S. Oreshin, A. Serdyukov [et al.] // ACM International Conference Proceeding Series, Xiamen, 27–29 October 2020. – Xiamen, 2020. – P. 84-87. – DOI 10.1145/3437802.3437817.

18.Implementing a Machine Learning Approach to Predicting Students Academic Outcomes / S. Oreshin, A. Filchenkov, P. Petrusha, ..., V. Kazakovtsev // ACM International Conference Proceeding Series, Xiamen, 27–29 октября 2020 года. – Xiamen, 2020. – P. 78-83. – DOI 10.1145/3437802.3437816.

19.Genetic Algorithms with the Crossover-Like Mutation Operator for the k-Means Problem / L. Kazakovtsev, G. Shkaberina, I. Rozhnov, R. Li, V. Kazakovtsev // Communications in Computer and Information Science. – 2020. – Vol. 1275. – P. 350-362. – DOI 10.1007/978-3-030-58657-7_28.

20.New method of training two-layer sigmoid neural networks using regularization / V. N. Krutikov, L. A. Kazakovtsev, G. Sh. Shkaberina, V. L. Kazakovtsev // IOP Conference Series: Materials Science and Engineering : International Workshop "Advanced Technologies in Material Science, Mechanical and Automation Engineering – MIP: Engineering – 2019", Krasnoyarsk, 04–06 апреля 2019 года. Vol. 537. – London: Institute of Physics and IOP Publishing Limited, 2019. – P. 42055.

21.Krutikov, V. N. Non-smooth regularization in radial artificial neural networks / V. N. Krutikov, L. A. Kazakovtsev, V. L. Kazakovtsev // IOP Conference Series: Materials Science and Engineering : Aerospace technologies, Krasnoyarsk, 20–28 октября 2018 года. – London: IOP science, 2018. – P. 042010. – DOI 10.1088/1757-899X/450/4/042010.

Прочие публикации

22.Методы автоматической группировки объектов в системах анализа и хранения данных : монография / Ф. Г. Ахматшин, Л. А. Казаковцев, В. Л. Казаковцев. – Москва : ООО "Научно-издательский центр Инфра-М", 2025. – 160 с. – ISBN 978-5-16-021831-1.

Свидетельства о регистрации ЭВМ:

23. Система интеллектуального анализа данных тестовых испытаний промышленной продукции на основе алгоритмов разделения смеси гауссовых распределений повышенной точности. Свидетельство о государственной регистрации программы для ЭВМ / Д. В. Сташков, Л. А. Казаковцев, Р. И. Кузьмич, В. Л. Казаковцев / Роспатент. рег. № 2017663875 от 13.02.2017: заявл. 16.10.2017.

24.Казаковцев В.Л. Multi Group Encoding. Свидетельство о государственной регистрации программы для ЭВМ / В.Л. Казаковцев, А.А. Ступина, Л.А. Казаковцев // Роспатент. рег. № 2026618245 от 24.03.2026. заявл. 17.02.2026.