

УТВЕРЖДАЮ

Проректор
по научной работе, к.ф.-м.н.



Баган Виталий
Анатольевич

«11» сентября 2025 г.

ЗАКЛЮЧЕНИЕ

федерального государственного автономного образовательного
учреждения высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
(МФТИ, Физтех)

Диссертация «Развитие метода динамичного формирования групп объектов по принципу идентичности для больших данных» выполнена на кафедре интеллектуальных систем федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» Министерства науки и высшего образования Российской Федерации.

В период подготовки диссертации с 2022 г. по настоящее время соискатель **Рябцев Антон Борисович** работал в компании Общество с ограниченной ответственностью «Озон Технологии» в отделе по продукту и технологиям «Машинное обучение и матчинг» в должности ведущего математика-разработчика.

В 2022 г. Рябцев А.Б. окончил федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)» по направлению подготовки 03.04.01 Прикладные математика и физика. В 2025 г. окончил аспирантуру МФТИ по программе подготовки научных и научно-педагогических кадров в аспирантуре «Информационные технологии и телекоммуникации».

Справка о сдаче кандидатских экзаменов выдана в 2025 г. федеральным государственным автономным образовательным учреждением высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

Научный руководитель – доктор технических наук, профессор Дулин Сергей Константинович. Основное место работы – федеральное государственное автономное образовательное учреждение высшего образования «Российский

университет транспорта», центр обеспечения и экспертиз научно-исследовательских проектов и программ развития транспортной отрасли Управления новых проектов и технологий, ведущий эксперт.

По итогам обсуждения диссертации «Развитие метода динамического формирования групп объектов по принципу идентичности для больших данных» **принято следующее заключение.**

Актуальность темы. В современных условиях объёмы обрабатываемой информации непрерывно растут, что приводит к усложнению процессов интеграции, сопоставления и анализа данных. Развитие технологий больших данных и распределённых вычислений делает особенно востребованными методы автоматической группировки схожих объектов. Грамотная группировка данных не только улучшает работу аналитических систем, но и способствует более эффективному принятию решений в различных сферах, включая бизнес, логистику и здравоохранение. Однако формирование групп усложняется в условиях высокой динамичности данных, когда характеристики объектов постоянно меняются, появляются новые элементы, а существующие удаляются. Традиционные алгоритмы кластеризации имеют существенные ограничения: они требуют заранее заданного числа кластеров, плохо масштабируются или чувствительны к параметрам. Это обуславливает необходимость разработки методов, способных работать с данными переменной структуры и адаптироваться к изменениям в реальном времени.

Отдельное внимание уделяется методам анализа динамических графов, где важно учитывать временные изменения, разрывы связей и слияние сообществ. Однако большинство существующих решений либо рассчитаны на слабую динамику, либо требуют значительных вычислительных ресурсов для постоянного обновления кластерной структуры. Ещё одной важной задачей является эффективный расчёт признаков для оценки идентичности объектов. Во многих прикладных областях такие признаки формируются на основе агрегации событийных данных с помощью SQL-запросов. Однако при работе с большими данными сложные аналитические запросы часто становятся узким местом. Современные исследования показывают, что применение машинного обучения для оптимизации SQL-запросов может значительно ускорить их выполнение, хотя внедрение таких методов связано с рядом технических сложностей.

Работа соискателя находится на стыке нескольких областей: теории структурной интероперабельности, динамической кластеризации и оптимизации аналитических вычислений. В ней предлагается усовершенствованный метод динамического формирования групп объектов на основе их идентичности, учитывающий требования реального времени, высокую изменчивость данных и необходимость поддержания согласованности групп.

Цели и задачи диссертации:

1. Провести комплексный анализ применения методов машинного обучения для оптимизации выполнения аналитических SQL-запросов в условиях изменяющихся данных и высоких нагрузок для определения возможности их применения в реальных системах обработки больших данных.

2. Выбрать архитектуру решения задачи формирования групп. Разработать новый алгоритм кластеризации для повышения структурной согласованности получаемых групп даже при наличии неполной или шумной информации.

3. Продемонстрировать возможность эффективной адаптации предложенного метода динамичной кластеризации для распределённых вычислительных систем, чтобы подчеркнуть его масштабируемость и применимость к обработке больших объёмов данных.

4. Разработать комплексную методику количественной оценки качества группировки объектов, включающую показатели однородности и полноты, адаптированные для анализа результатов динамичной кластеризации в условиях больших данных.

Основные результаты диссертации:

1. Проведён анализ модифицированного подхода к оптимизации SQL-запросов с использованием машинного обучения. Установлены ключевые проблемы применения ML-моделей в реальных условиях: сложность адаптации к изменяющимся данным и нагрузке и высокие вычислительные затраты на поддержание актуальности моделей при динамическом изменении структуры данных.

2. Разработан метод повышения качества данных за счёт идентификации семантически идентичных объектов. Предложена гибридная архитектура, включающая модель бинарной классификации для оценки схожести объектов и алгоритм распространения меток (LPA) для группировки. Показана возможность масштабирования подхода в распределённых системах (на основе Hadoop). Подтверждена эффективность метода на реальных данных маркетплейсов (задача формирования групп идентичных товаров).

3. Предложен усовершенствованный двухэтапный алгоритм кластеризации на основе LPA. Обосновано преимущество LPA для группировки семантически схожих объектов. Разработана модификация LPA с адаптивной калибровкой параметров. Эксперименты подтвердили высокую эффективность метода в задачах управления ассортиментом при работе с большими динамическими данными.

4. Разработана методика регулярной оценки качества группировки объектов в производственных системах: разработан способ оценки однородности и полноты в условиях больших данных.

Все ***результаты диссертации получены лично соискателем*** при научном руководстве доктора технических наук, профессора Дулина С.К.

Научная новизна работы. Впервые проведён комплексный анализ применения методов машинного обучения для оптимизации аналитических SQL-запросов в условиях изменяющихся данных и высоких нагрузок, выявивший ключевые ограничения их использования в реальных системах обработки больших данных. Обосновано применение комбинированного подхода, сочетающего бинарную классификацию для оценки парной схожести объектов и алгоритм распространения меток (LPA) для формирования групп, что обеспечивает эффективное решение задачи динамического объединения семантически идентичных объектов при ограниченных вычислительных

ресурсах. Разработан модифицированный двухэтапный алгоритм кластеризации на основе LPA с адаптивной калибровкой пороговых параметров, повышающий структурную согласованность групп даже при наличии неполных или зашумлённых данных. Доказана возможность масштабирования предложенного метода для распределённых систем, что расширяет его применимость к обработке больших объёмов информации. Кроме того, создана комплексная методика оценки качества группировки, включающая специализированные метрики однородности и полноты, адаптированные для анализа результатов динамической кластеризации в условиях больших данных.

Практическая ценность. Разработанный подход внедрён в систему формирования групп идентичных товарных предложений в компании ООО «Озон Технологии». Качество работы системы ежедневно оценивается с помощью предложенной методики. Результаты работы системы используются для формирования блоков «Предложения других продавцов», «Есть дешевле» и «Есть более быстрая доставка».

Обоснованность и достоверность результатов и выводов

Выводы диссертации обоснованы аналитическими расчетами и подтверждены данными лабораторных и научных экспериментов, полученными с помощью разработанных алгоритмов и комплексов программ. Теоретическую и методологическую основу проведенных разработок и исследований составили труды отечественных и зарубежных авторов в области анализа данных и СУБД. Для анализа полученных результатов использовались методы статистической обработки данных. Достоверность результатов исследования обеспечена внедрением разработанных методов и алгоритмов в производственную систему крупнейшего маркетплейса страны, что подтверждает их практическую эффективность и доказанную экономическую пользу, а также апробацией предложенных решений на больших и представительных массивах данных.

Материалы диссертации опубликованы соискателем достаточно полно в следующих работах:

1. Дулин С.К., Рябцев А.Б. Анализ подходов к оптимизации запросов в аналитических СУБД // Образовательные ресурсы и технологии. – 2023. – № 3 (44). – С. 73–80.
2. Дулин С.К., Рябцев А.Б. Алгоритм улучшения согласованности структурной интероперабельности // Надёжность. – 2024. – Т. 24, № 2. – С. 8–15.
3. Антипов И.Ф., Дулин С.К., Рябцев А.Б. Формирование групп идентичных объектов // Известия РАН. Теория и системы управления. – 2025. – № 3. – С. 113–120.
4. Дулин С.К., Рябцев А.Б. Оценка планов выполнения SQL запросов для решения транспортных задач // Наука и технологии железных дорог. – 2023. – Т. 7, № 1 (25). – С. 38–43.

5. Дулин С.К., Рябцев А.Б. Анализ концептов для проектирования базы геоданных железнодорожных геоописаний // Наука и технологии железных дорог. – 2024. – Т. 8, № 3 (31). – С. 24–32.

Личный вклад соискателя в работах с соавторами заключается в следующем:

[1,5] – реализация методов, проведение численных экспериментов, анализ результатов; [2] – реализация алгоритма, сравнение с другим алгоритмом; [3] – разработка алгоритма, проведение численных экспериментов, анализ результатов; [4] – анализ результатов.

Основные результаты работы докладывались на следующих научных конференциях и семинарах:

1. 14-я Международная конференция «Интеллектуализация обработки информации» (ИОИ-2022), Москва, 2022.
2. 66-я Всероссийская научная конференция МФТИ, Долгопрудный, 2024.
3. 15-я Международная конференция «Интеллектуализация обработки информации» (ИОИ-2024), Гродно, 2024.

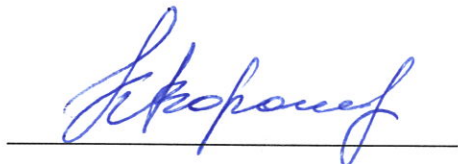
Содержание диссертации соответствует паспорту специальности
2.3.8. Информатика и информационные процессы, в частности, пунктам:

- пункт 1: Разработка компьютерных методов и моделей описания, оценки и оптимизации информационных процессов и ресурсов, а также средств анализа и выявления закономерностей на основе обмена информацией пользователями и возможностей используемого программно-аппаратного обеспечения.
- пункт 7: Разработка методов обработки, группировки и аннотирования информации, в том числе, извлеченной из сети интернет, для систем поддержки принятия решений, интеллектуального поиска, анализа.
- пункт 12: Разработка технологий извлечения и анализа информации в больших базах данных, в том числе, с использованием концепции многомерного представления (OLAP) и интеллектуального анализа данных (Data Mining) статического и в реальном масштабе времени, реализация моделей баз знаний.
- пункт 13: Разработка и применение методов распознавания образов, кластерного анализа, нейро-сетевых и нечетких технологий, решающих правил, мягких вычислений при анализе разнородной информации в базах данных.

Диссертация соискателя Рябцева Антона Борисовича «Развитие метода динамичного формирования групп объектов по принципу идентичности для больших данных» является законченной научно-квалификационной работой, которая соответствует требованиям пунктов 9, 10 и 14 Положения о присуждении учёных степеней, утверждённого постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842.

Диссертация «Развитие метода динамического формирования групп объектов по принципу идентичности для больших данных» **Рябцева Антона Борисовича** рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности **2.3.8 Информатика и информационные процессы**.

Заключение принято на заседании кафедры Интеллектуальных систем МФТИ, Физтех. Присутствовало на заседании 10 человек. Результаты голосования: «за» – 10 чел., «против» – нет, «воздержались» – нет, протокол №7 от 11 сентября 2025 г.



Воронцов Константин Вячеславович,
д-р физ.-мат. наук, профессор
РАН, заведующий кафедрой
интеллектуальных систем МФТИ