На правах рукописи

Рябцев Антон Борисович

РАЗВИТИЕ МЕТОДА ДИНАМИЧНОГО ФОРМИРОВАНИЯ ГРУПП ОБЪЕКТОВ ПО ПРИНЦИПУ ИДЕНТИЧНОСТИ ДЛЯ БОЛЬШИХ ДАННЫХ

Специальность 2.3.8 —
«Информатика и информационные процессы (технические науки)»

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Москва — 2025

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

Научный руководитель: **Дулин Сергей Константинович,**
доктор технических наук, профессор,
Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет транспорта»,
ведущий эксперт

Официальные оппоненты: **Миркин Борис Григорьевич,**
доктор технических наук,
профессор департамента анализа данных и искусственного интеллекта Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Голосов Павел Евгеньевич,
кандидат технических наук,
директор Института общественных наук Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации

Ведущая организация: Акционерное общество «Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте»

Защита состоится « ____ » _____ 2025 г. в ___:___ на заседании диссертационного совета 24.1.224.03 на базе Федерального исследовательского центра «Информатика и управление» Российской академии наук по адресу: 119333, Москва, ул. Вавилова, д.42.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук и на сайте http://www.frccsc.ru.

Автореферат разослан _____ 2025 г.

Ученый секретарь
диссертационного совета
24.1.224.03, к.т.н. Рейер И.А.

Общая характеристика работы

Актуальность темы. В последние десятилетия наблюдается стремительный рост объёмов обрабатываемых данных, что сопровождается усложнением задач их интеграции, сопоставления и анализа. На фоне развития технологий больших данных, распределённых вычислений и методов искусственного интеллекта одной из ключевых задач становится эффективное объединение идентичных или схожих объектов в группы. Корректная группировка позволяет не только повысить качество аналитических систем, но и существенно оптимизировать процессы принятия решений в бизнесе, логистике, здравоохранении и других отраслях.

Формирование групп идентичных объектов особенно усложняется в условиях высокой динамичности данных. Постоянные изменения в характеристиках объектов, появление новых и удаление уже имеющихся требуют от систем быстрой адаптации и переоценки связей между элементами. При этом классические методы кластеризации, такие как K-means, требуют заранее заданного числа кластеров и поэтому не могут быть непосредственно применены к задаче динамичного формирования групп. Методы на основе плотности, такие как DBSCAN, также имеют ограничения по масштабируемости и чувствительности к выбору параметров. Это приводит к необходимости выбора метода, который способен эффективно обрабатывать данные с заранее неизвестной структурой и числом групп, и его развитие для успешного применения в динамичных системах. Одним из активно развивающихся направлений является исследование методов повышения структурной согласованности в динамичных графах. Современные обзоры показывают, что для эффективного обнаружения сообществ в меняющихся сетях необходимы подходы, способные учитывать временные аспекты, разрывы связей и слияния сообществ. Тем не менее, подавляющее большинство существующих решений ориентировано на слабую динамику или требуют значительных ресурсов для регулярного пересчёта кластерной структуры. Помимо задач структурного группирования, критически важной становится проблема быстрого и эффективного расчёта признаков для оценки идентичности объектов. Во многих прикладных задачах такие признаки строятся на основе агрегатов событийных данных, извлекаемых с помощью аналитических SQL-запросов. Однако выполнение сложных аналитических запросов в условиях больших данных зачастую становится узким

местом. Современные исследования демонстрируют, что применение методов машинного обучения в оптимизаторах SQL-запросов позволяет добиться существенного повышения их производительности, хотя внедрение подобных решений сопряжено с рядом технологических и практических трудностей.

Таким образом, представленное исследование находится на пересечении нескольких направлений — подхода к улучшению интероперабельности, динамичной кластеризации и оптимизации аналитических вычислений. В работе предлагается развитие метода динамичного формирования групп объектов по принципу идентичности с учётом ограничений реального времени, высокой изменчивости системы и необходимости поддержания высокой согласованности групп.

Целью данной работы является развитие метода динамичного формирования групп объектов по принципу идентичности в условиях больших и изменчивых данных, обеспечивающего высокую структурную согласованность групп при учёте необходимости быстрого вычисления признаков идентичности.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Анализ существующих методов группировки объектов и выявление их ограничений при работе с большими динамичными данными.
2. Исследование возможностей ускорения аналитических вычислений, необходимых для построения признаков объектов, с применением методов машинного обучения для оптимизации выполнения SQL-запросов.
3. Разработка алгоритма формирования групп объектов, устойчивого к ошибкам в определении идентичности и способного поддерживать согласованную структуру в условиях высокой динамики данных.
4. Создание методики оценки качества сформированных групп с точки зрения однородности и полноты разбиения, адаптированной для работы с большими и быстро изменяющимися данными.

Таким образом, работа направлена на комплексное решение задачи структурно согласованного формирования групп идентичных объектов с учётом особенностей выполнения вычислений в больших динамичных системах.

Научная новизна:

1. Впервые проведён комплексный анализ применения методов машинного обучения для оптимизации выполнения аналитических SQL-запросов в условиях изменяющихся данных и высоких нагрузок, что позволило выявить ограничения по их практической применимости в реальных системах обработки больших данных.
2. Обоснован выбор сочетания бинарной классификации для оценки парной схожести объектов и алгоритма распространения меток (Label Propagation Algorithm) для формирования групп как наиболее эффективной комбинации в задачах динамичного объединения семантически идентичных объектов при ограниченных вычислительных ресурсах.
3. Предложен модифицированный двухэтапный алгоритм кластеризации на основе LPA с калибровкой пороговых параметров на основе допустимой доли ошибочных связей (5% и 20%), что позволяет повысить структурную согласованность получаемых групп даже при наличии неполной или шумной информации.
4. Доказана возможность эффективной адаптации предложенного метода динамичной кластеризации для распределённых вычислительных систем, что обеспечивает его масштабируемость и применимость к обработке больших объёмов данных.
5. Разработана комплексная методика количественной оценки качества группировки объектов, включающая показатели однородности и полноты, адаптированные для анализа результатов динамичной кластеризации в условиях больших данных.

Научная и практическая значимость. Теоретическая значимость работы заключается в развитии подходов к динамичному формированию групп объектов по принципу идентичности в условиях больших и изменяющихся данных. В работе обоснован выбор комбинации бинарной классификации и алгоритма распространения меток (LPA) как эффективного решения для задач динамичной кластеризации при ограниченных вычислительных ресурсах. Также предложена модификация процесса кластеризации, обеспечивающая повышение структурной согласованности групп за счёт калибровки пороговых параметров для установки допустимой доли ошибок. Разработанная методика оцен-

ки качества группировки объектов, учитывающая показатели однородности и полноты, расширяет инструментарий анализа динамичных кластеризаций и может быть использована в других задачах обработки больших данных.

Практическая значимость работы заключается в возможности применения разработанных методов и алгоритмов для широкого круга прикладных задач, связанных с обработкой больших динамичных данных. К таким задачам относятся: автоматизированное управление ассортиментом товаров, интеллектуальная агрегация предложений на маркетплейсах, анализ и сопоставление записей в базах данных, системы мониторинга состояния объектов, управление цифровыми двойниками, задачи интеграции данных из разнородных источников, очистка данных от дубликатов и повышение их качества. Разработанный подход адаптирован для распределённых вычислительных систем, что обеспечивает его масштабируемость и позволяет эффективно работать с большими объёмами данных. Проведённые эксперименты подтверждают практическую эффективность предложенных решений на реальных задачах, демонстрируя высокое качество группировки объектов и устойчивость методов к изменениям данных. Результаты диссертационного исследования использованы при реализации внутренних проектов отдела по продукту и технологиям «Машинное обучение и матчинг» ООО «Озон Технологии» в 2021-2025 годах.

Основные положения, выносимые на защиту:

1. Проанализирован модифицированный подход на основе машинного обучения для оптимизации аналитических SQL-запросов в СУБД:
    (a) Выявлена сложность адаптации моделей машинного обучения к изменяющимся данным и нагрузкам в условиях реального времени.
    (b) Выявлены высокие накладные расходы на поддержание актуальности моделей при динамичной смене структуры данных.
2. Предложен практико-ориентированный метод повышения качества данных через идентификацию семантически идентичных объектов:
    (a) Предложена комбинированная архитектура, сочетающая модель бинарной классификации для оценки степени схожести объектов и алгоритм распространения меток (LPA) для последующего формирования групп.

(b) Обоснована возможность адаптации разработанного подхода для распределённых вычислительных систем с использованием парадигмы MapReduce.

(c) Доказана практическая эффективность предложенного метода на задаче поиска идентичных товарных предложений на маркетплейсе.

3. Предложен модифицированный двухэтапный алгоритм кластеризации на основе LPA:

   (a) Обоснован выбор алгоритма LPA как оптимального решения для задачи формирования групп семантически идентичных объектов.

   (b) Предложена оригинальная двухэтапная модификация LPA, использующая последовательное построение сообществ на основе графов с высокой и средней надёжностью связей между объектами.

   (c) Экспериментально подтверждена высокая эффективность предложенного метода при решении практических задач управления ассортиментом в условиях больших динамичных данных.

4. Предложена комплексная методика оценки качества группировки объектов, адаптированная для работы с большими и быстро изменяющимися данными:

   (a) Предложен метод оценки однородности кластеров как меры внутреннего качества группировки.

   (b) Предложен метод оценки полноты группировки как меры соответствия найденных групп истинной структуре данных.

Достоверность результатов обеспечивается обширным анализом работ в области исследования, описанием проведённых экспериментов, их воспроизводимостью, апробацией результатов на практике. Основные результаты диссертации докладывались на следующих конференциях: 14-я Международная конференция «Интеллектуализация обработки информации» (ИОИ-2022), Москва, 2022; 66-я Всероссийская научная конференция МФТИ, Долгопрудный, 2024; 15-я Международная конференция «Интеллектуализация обработки

информации» (ИОИ-2024), Гродно, 2024. Результаты данной работы использованы в рамках государственного задания номер 103-00001-25-02.

Публикации. Материалы диссертации опубликованы в 8 печатных работах, из них 3 в журналах из списка ВАК.

Личный вклад. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Подготовка к публикации полученных результатов проводилась совместно с соавторами, причём вклад диссертанта был определяющим. Все представленные в диссертации результаты получены лично автором.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава диссертационной работы посвящена изложению теоретических основ задачи динамичного формирования групп объектов по принципу идентичности, с акцентом на понятиях интероперабельности, структурной согласованности и оценки качества группировки.

Глава начинается с анализа понятия интероперабельности — способности различных информационных систем и компонентов обмениваться данными, интерпретировать их и использовать для согласованных действий. Рассматриваются её ключевые аспекты: техническая, семантическая и организационная интероперабельность (рис. 1). Особое внимание уделено структурной интероперабельности, которая играет решающую роль при организации взаимодействия между объектами в динамичных и распределённых системах, где важна не только корректная передача данных, но и согласованность их структуры и связей. Показано, что структурная интероперабельность выходит за рамки форматов и протоколов обмена и требует устойчивой и логически непротиворечивой модели связей между элементами системы.

Вторая часть главы посвящена понятию структурной согласованности — степени соответствия фактической структуры связей между объектами некоторой идеальной модели. В частности, рассматриваются требования к ”консо-

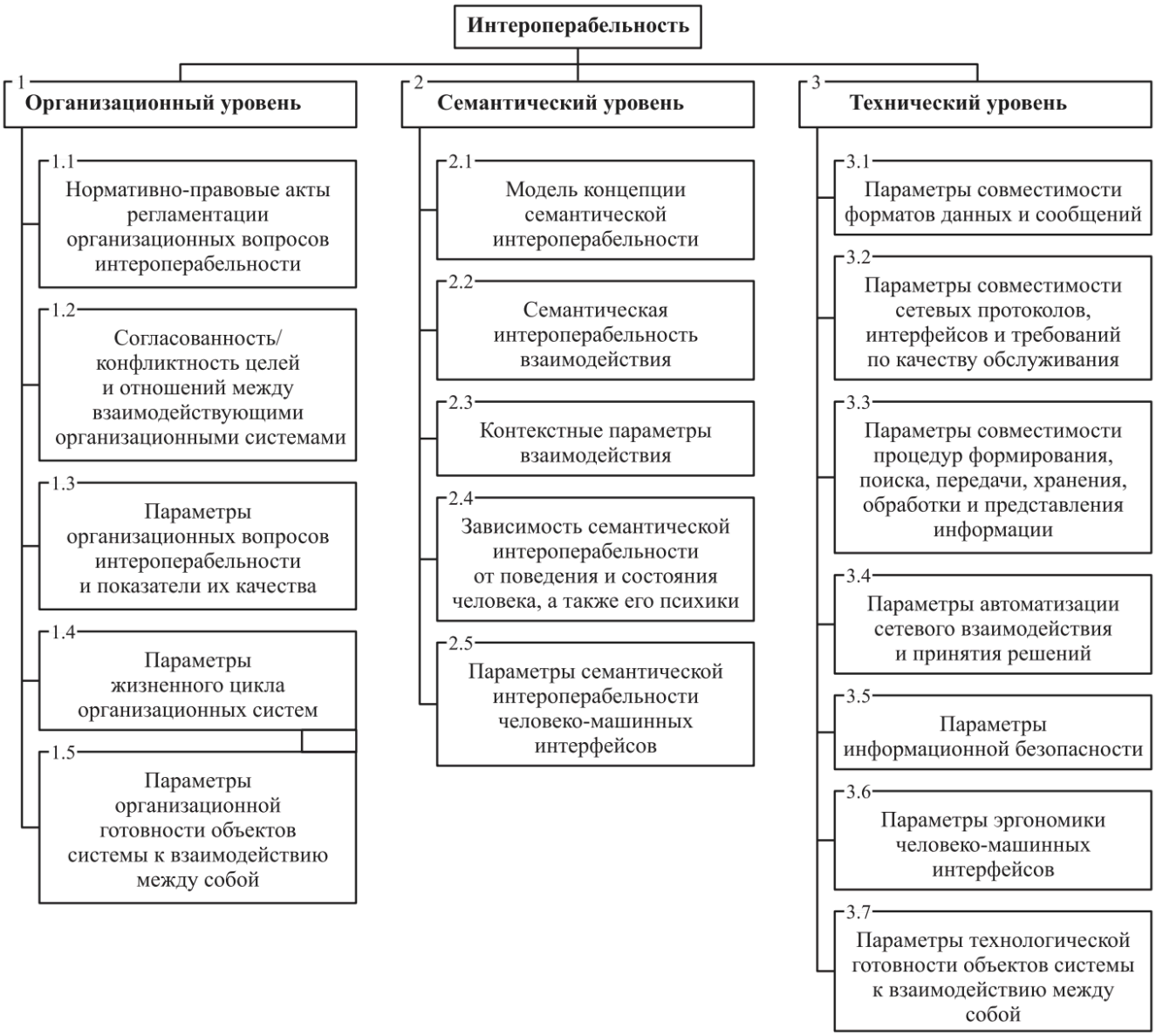

Рис. 1 — Общая структура интероперабельности в соответствии с ГОСТ Р 55062-2012

нансной” структуре (без межгрупповых связей и с высокой плотностью внутригрупповых), а также описываются типичные отклонения от идеала в реальных данных, ведущие к так называемым ”ассонансным” структурам. Обсуждается, каким образом шум, ошибки, пропуски и динамика данных приводят к нарушению согласованности и затрудняют формирование корректных групп. Структурная согласованность в этом контексте выступает не как бинарное свойство, а как количественно измеримая характеристика, подлежащая оценке и оптимизации.

Следующий раздел главы содержит обзор методов оценки согласованности структуры. Подробно рассматриваются такие меры качества, как однородность, полнота, V-мера, индекс Рэнда с поправкой на случайность (ARI), оценка

Фаулкса–Мэллоуза и другие. Каждая из них позволяет измерять качество полученной кластерной структуры с разных точек зрения: насколько хорошо группировка объединяет действительно схожие объекты, не теряя при этом охвата. Отдельно подчёркивается, что в задачах с динамично изменяющимися данными критически важным становится анализ устойчивости группировки во времени и способность адаптивно поддерживать согласованность при эволюции входных данных.

В завершающей части главы проводится формализация задачи динамичного формирования групп объектов по принципу идентичности. Обозначаются основные требования к алгоритмам: отсутствие необходимости заранее знать число групп, способность работать с большими объёмами данных, устойчивость к ошибкам и адаптивность к изменениям. Даётся строгое математическое описание задачи, включающее множества объектов, их признаков, граф связей между ними, и целевое разбиение на устойчивые кластеры. Формулируются критерии, которым должно удовлетворять решение задачи: высокая однородность и полнота, масштабируемость, устойчивость к шуму и возможность инкрементального обновления.

В рамках поставленной задачи в работе обосновывается выбор подхода, сочетающего бинарную модель оценки идентичности объектов и графовые алгоритмы кластеризации. В частности, в качестве базового алгоритма группировки выбирается Label Propagation Algorithm (LPA), что закладывает основу для последующих разработок и экспериментов, представленных в последующих главах диссертации.

Вторая глава диссертации посвящена задаче оценки схожести объектов как ключевому этапу, предшествующему их группировке по принципу идентичности. Сама возможность адекватного объединения объектов в группы предполагает наличие механизма сравнения, позволяющего количественно выразить степень их схожести.

В первой части главы формализуется общая постановка задачи: имеются объекты, каждый из которых описывается набором признаков, и необходимо определить, насколько два объекта «похожи» в семантическом смысле, чтобы считать их идентичными с точки зрения реальной сущности. В условиях отсутствия универсальных идентификаторов и высокого уровня неоднородности

данных этот процесс требует использования приближённых методов, основанных на вычислении меры схожести.

Функция схожести определяется как отображение, принимающее пару объектов и возвращающее значение из интервала $[0,1]$, отражающее степень уверенности в их идентичности. В работе приводится один из возможных вариантов такой функции, реализуемой через агрегирование отклонений по признакам с использованием весов и нормализаций. Этот пример демонстрирует, каким образом на практике может быть сконструирована оценка, но не является универсальным определением для всех случаев. При этом подчёркивается, что структура признаков, их шкалы, наличие пропусков и разная степень важности сильно влияют на форму и поведение функции.

На основе значений функции схожести между всеми парами объектов формируется матрица, представляющая собой частичное отображение на подмножестве пар. В типичных сценариях реальных данных значительная часть парных сравнений либо невозможна (например, по отсутствующим данным), либо их схожесть заведомо близка к нулю, поэтому соответствующая матрица оказывается разреженной. Из неё может быть построен граф, в котором рёбра отражают наличие значимой связи между парой объектов, если значение функции схожести превышает выбранный порог.

Порог значения схожести — важный параметр, определяющий структуру получаемого графа. При высоких значениях порога повышается Precision, но страдает Recall. При низком пороге ситуация обратная: растёт число ошибочных связей между разными сущностями. Таким образом, выбор порога всегда представляет собой компромисс, который должен соотноситься с целями системы.

Важной характеристикой модели схожести являются типы ошибок, которые она допускает. Основными из них являются ложноположительные (false positives) и ложноотрицательные (false negatives). Первые приводят к неправильному объединению различных объектов, вторые — к фрагментации одной сущности на несколько кластеров. В графовой постановке задачи это означает, соответственно, появление рёбер между объектами разных групп или отсутствие рёбер внутри одной группы. Эти ошибки напрямую влияют на структуру кластеров и общее качество результирующего разбиения.

Особый интерес в работе уделяется практическим аспектам формирования признаков, необходимых для работы функции схожести. В ряде прикладных сценариев признаки объектов не присутствуют в данных явно, а выводятся из накопленных событий или поведения. Например, в задачах, связанных с e-commerce, логистикой, пользовательскими профилями, телеметрией или регистрами, признаки формируются как агрегаты по истории: количество операций, разнообразие характеристик, статистики за временные окна и т.п.

Получение таких признаков требует выполнения сложных аналитических запросов, зачастую — многократно, при каждом изменении объекта. Это приводит к серьёзным вычислительным издержкам: высокие задержки, нагрузка на хранилища, проблемы синхронизации. В условиях динамики, когда объекты изменяются постоянно, признаковая модель должна быть переоценена, что делает применение ресурсоёмких моделей затруднительным. В работе подчёркивается, что даже при наличии высокоточной функции схожести, эффективность всей системы может быть ограничена скоростью и стабильностью извлечения признаков.

В завершение главы обсуждаются архитектурные ограничения и возможные пути их частичного смягчения — через кэширование, материализацию подзапросов, асинхронные вычисления, периодическое обновление признаков и их декомпозицию. Однако в работе делается вывод, что эти меры не устраняют корневую проблему — высокий вычислительный порог. Вследствие этого ставится задача анализа и оптимизации выполнения аналитических запросов, что и становится предметом следующей главы.

Третья глава диссертации посвящена исследованию одного из центральных технических ограничений всей системы формирования групп — медленного выполнения аналитических SQL-запросов, необходимых для построения признаков объектов. Преобразование данных в высокоуровневые признаки — неотъемлемая часть моделирования идентичности объектов, и узкое место здесь — это именно производительность запросов, особенно в условиях распределённых хранилищ и массово-параллельных СУБД.

Глава содержит три крупных раздела: анализ традиционных методов оптимизации запросов, рассмотрение современных подходов на основе машинного обучения и вывод о пределах применимости исследуемых решений.

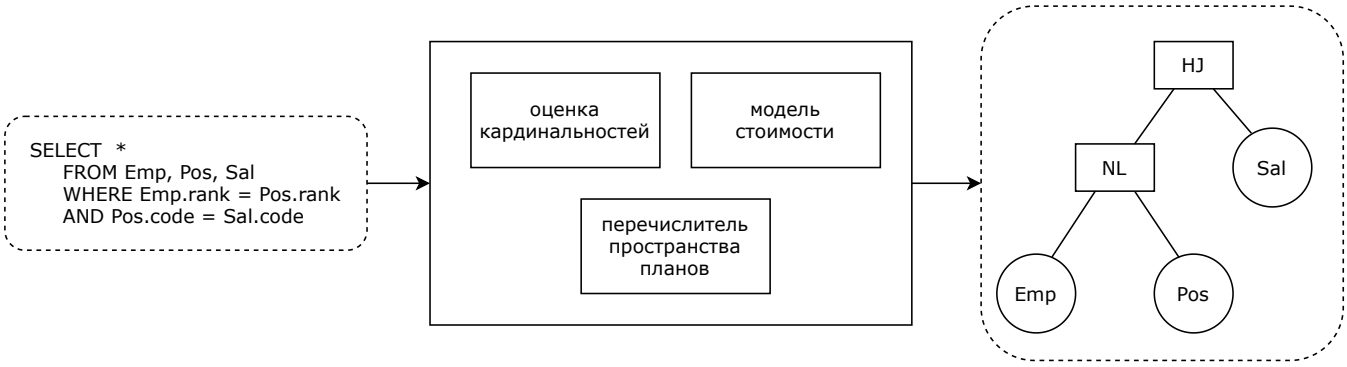

Рис. 2 — Архитектура традиционного оптимизатора запросов.

Первый раздел главы посвящён традиционным подходам к оптимизации SQL-запросов. Рассматривается архитектура классического оптимизатора, включающая подсистемы оценки кардинальности, модель стоимости и перечислитель планов (рис. 2). Показано, что эффективность всей системы зависит от точности оценки объёмов промежуточных данных — ключевого параметра, используемого при планировании выполнения запроса. Приводятся особенности реализации оптимизации в системах PostgreSQL и GaussDB, включая поддерживаемые стратегии соединения (Nested Loop, Merge Join, Hash Join) и их применимость к аналитическим нагрузкам.

Описывается роль модуля оценки кардинальности (CardEst), традиционно основанного либо на гистограммах, либо на сэмплировании. Указывается, что в реальных сценариях предположения о независимости и равномерности данных, лежащие в основе этих подходов, нередко нарушаются, что приводит к ошибочным оценкам и неэффективным планам.

Приводится обзор различий архитектурных решений для PostgreSQL и GaussDB, где в последней, как в СУБД массово-параллельной архитектуры, используется только Hash Join, а для обеспечения выполнения соединений требуется перераспределение данных — Broadcast или Redistribute, что влечёт за собой дополнительные затраты.

Второй раздел посвящён современным подходам к повышению эффективности выполнения SQL-запросов. Наиболее подробно анализируются методы, использующие машинное обучение как для улучшения оценки кардинальности, так и для непосредственного выбора плана выполнения.

В частности, рассматриваются:

– AQO (Adaptive Query Optimizer) — простой и доступный модуль в PostgreSQL, основанный на подстройке коэффициентов в форму-

ле селективности. В работе проведено тестирование AQO в OLAP-нагрузке, показано отсутствие ускорения.

– Naru и NeuroCard — модели на основе авторегрессии, обучаемые без выполнения запросов. Установлено, что они имеют ограниченную применимость в промышленных условиях из-за экспоненциального роста числа моделей при увеличении числа таблиц.
– PessimisticCardEst и FLAT — подходы, предлагающие альтернативную постановку задачи оценки кардинальности как оценки сверху или декомпозиции данных по связности. Отмечено, что практическое внедрение этих методов ограничено отсутствием кода, сложностью реализации и слабыми результатами при циклических соединениях.

Сравнительный анализ показывает, что улучшение оценки кардинальности не всегда приводит к ускорению выполнения запросов (рис. 3). Это связано с тем, что современные оптимизаторы зависят от собственных стоимостных моделей, и даже точная кардинальность не гарантирует выбора эффективного плана.

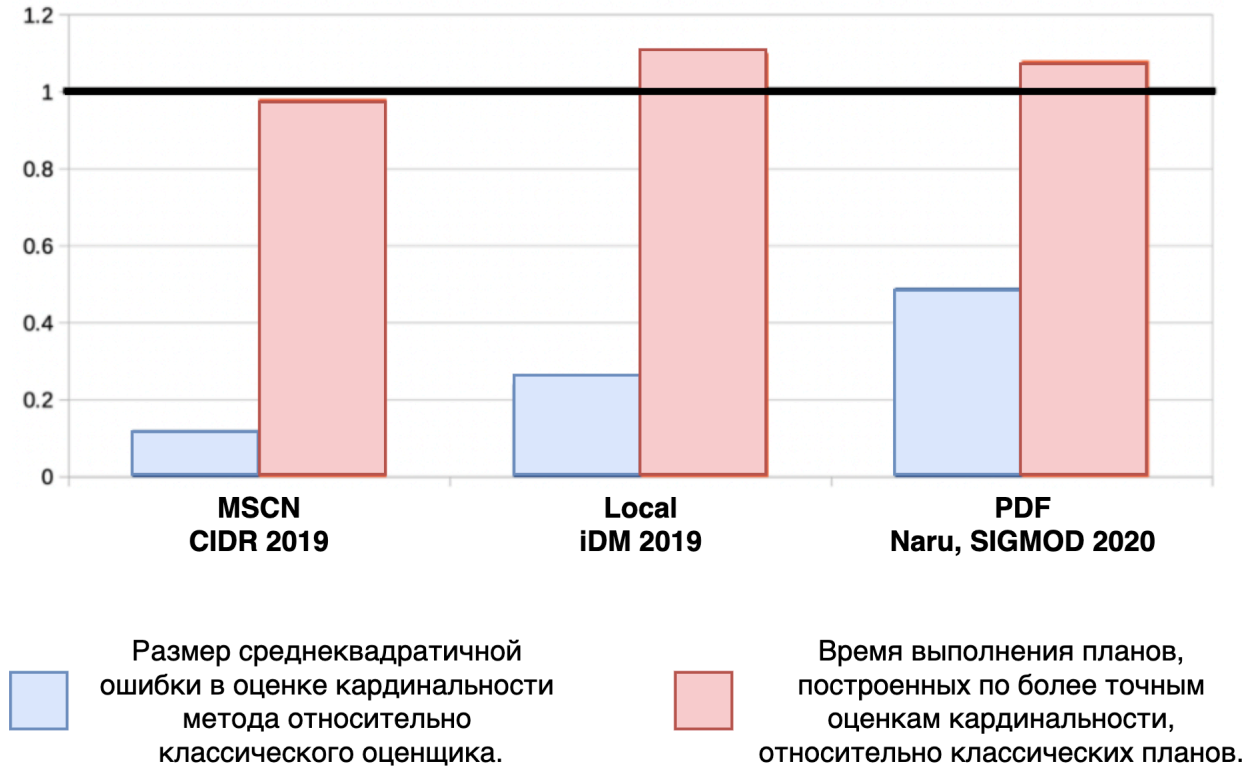

Рис. 3 — Сопоставление точности оценок кардинальности разными методами и времени выполнения запросов.

Поэтому особое внимание в работе уделено альтернативному направлению — прямой оптимизации плана запроса, минуя промежуточный этап предсказания кардинальности.

Важную часть главы составляет анализ подходов, основанных на обучении с подкреплением, прежде всего алгоритма DQN (Deep Q-Network). В диссертации подробно описана формализация задачи планирования соединений в виде марковского процесса принятия решений (MDP) (рис. 4). Это позволяет рассматривать оптимизацию плана как поиск наилучшей последовательности действий в пространстве возможных соединений, с учётом их стоимости и долгосрочных последствий.

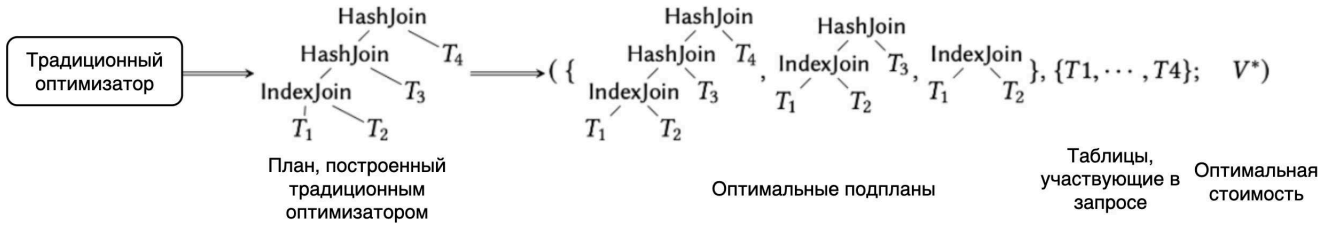

Рис. 4 — Используя принцип оптимальности, из одного плана, созданного собственным оптимизатором, извлекаются три обучающих примера. Эти примеры имеют одни и те же долгосрочные затраты и отношения для соединения (т.е. принятие этих локальных решений в конечном итоге приводит к соединению в одну связную компоненту $\{T1, ...,T4\}$ с оптимальной совокупной стоимостью V^*).

Показано, что использование DQN позволяет переобучать модель на результатах предыдущих планов и учитывать ошибки оценок, приводящих к неэффективным решениям. На вход модели подаются векторные признаки графа запроса, описания соединений, предикатов, типов операторов и частичных планов в виде деревьев признаков (рис. 5, 6). Обсуждаются подходы к кодированию входных данных, особенности функции потерь и стратегия генерации обучающих примеров из истории выполнения оптимизатора.

Наиболее проработанным подходом в этой области является Neo (Neural Query Optimizer). Его архитектура полностью реализует обучение модели оценки стоимости запроса на базе эмпирических данных, без использования вручную заданной стоимостной функции. Neo объединяет планировщик, оценщик кардинальности и модель стоимости в единую нейросетевую систему. Подробно анализируется архитектура модели, способы кодирования информации о запросе и плане, механизм сбора опыта и переобучения (рис. 7).

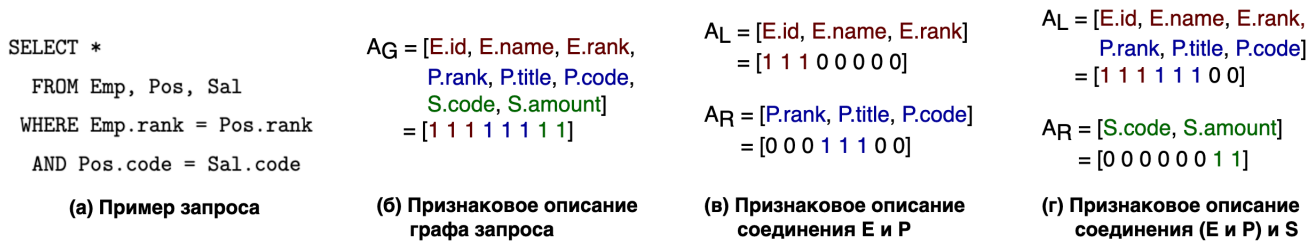

Рис. 5 — Запрос и соответствующее ему признаковое описание. Бинарные векторы кодируют атрибуты в графе запроса (A_G), левой части соединения (A_L) и правой части (A_R). Такое кодирование позволяет описать как граф запроса, так и конкретное соединение. Показаны промежуточное соединение и финальное соединение. Пример запроса охватывает все отношения в схеме, поэтому $A_G = A$.

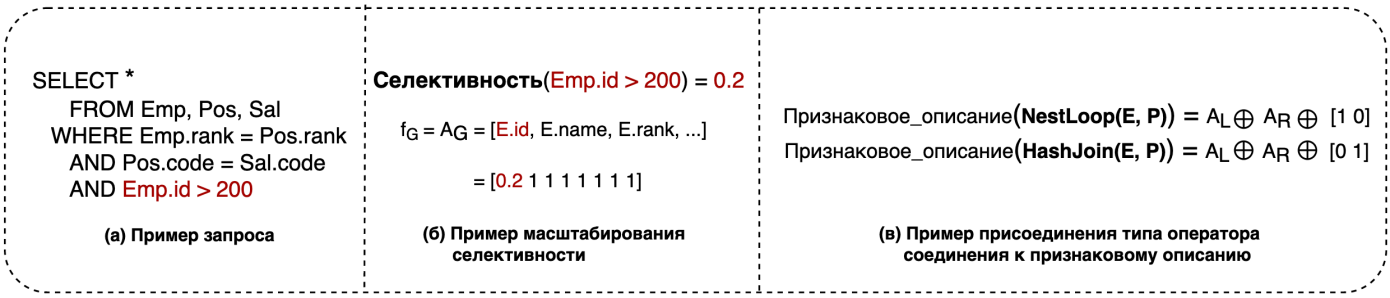

Рис. 6 — Учет предикатов и физических операторов (операторов соединения). Простые изменения базовой формы признакового описания необходимы для поддержки предикатов (слева) и физических операторов (справа). Например, если предположить, что система выбирает только между NestLoop и HashJoin, двумерный бинарный вектор объединяется с каждым вектором признаков соединения.

В заключение главы делается вывод о пределах применимости рассмотренных методов. Несмотря на наличие моделей, показывающих ускорение выполнения на тестовых данных (таблица 1), показано, что в реальных условиях, при изменении базы данных, такие модели требуют переобучения, что само по себе является дорогостоящей задачей. Более того, создание системы, которая бы отслеживала актуальность модели и принимала решение о её обновлении, представляет собой отдельную сложную инженерную задачу.

	Время выполнения 100 запросов.			
	Табличный режим СУБД		**Колоночный режим СУБД**	
	Традиционный оптимизатор	**“Умный” оптимизатор**	**Традиционный оптимизатор**	**“Умный” оптимизатор**
Вариант из оригинальной статьи.	**2.5 часа**	**5 часов**	**1.5 часа**	**3 часа**
Вариант с рядом модификаций.		**1.5 часа**		**2 часа**

Таблица 1 — Анализ результатов подхода Neo

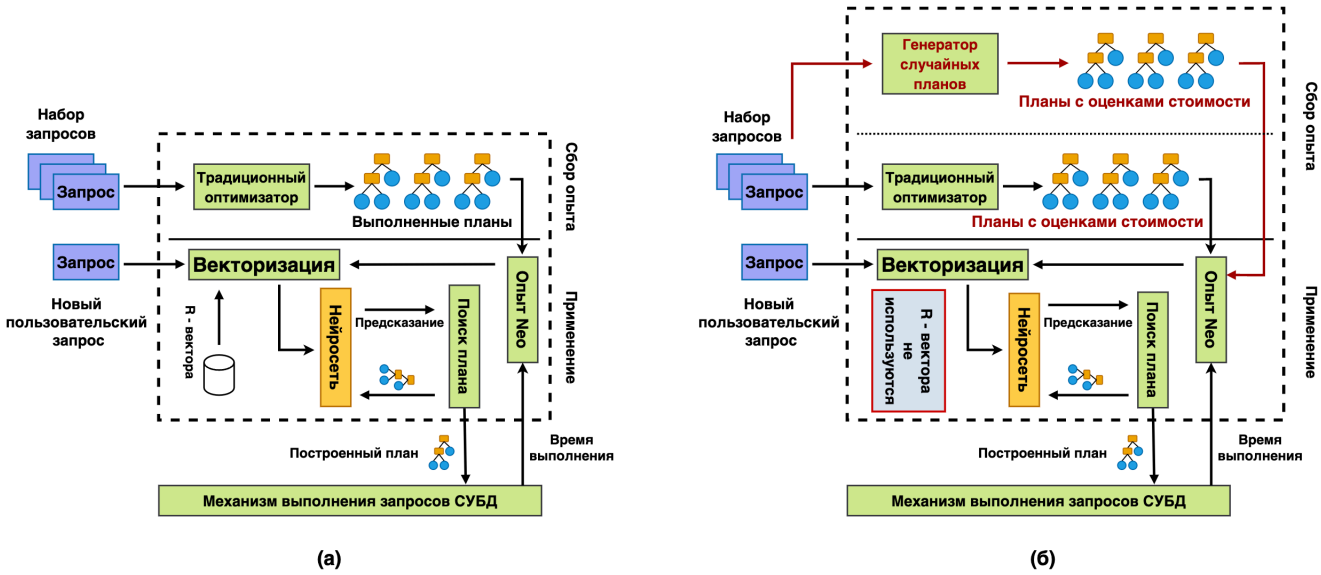

Рис. 7 —
(а) Дизайн системы Neo в оригинале.
(б) Модифицированный дизайн системы Neo – модификации подсвечены бордовым цветом.

Также фиксируется, что единственный подход, показавший реальное ускорение — Neo — требует значительных вычислительных ресурсов и архитектурных изменений, поэтому его использование на практике ограничено. В итоге делается вывод, что для задач, где признаки объектов требуют выполнения сложных SQL-запросов, универсального решения пока не существует. Однако существует широкий класс задач, в которых признаки либо заданы явно, либо извлекаются быстрее, и работа переключается на них — к исследованию методов группировки объектов по принципу идентичности.

Четвертая глава диссертации посвящена развитию и экспериментальному обоснованию выбора метода динамичного формирования групп объектов по принципу идентичности на основе графовой модели. В главе последовательно рассматриваются: формализация структуры связей между объектами, сравнительный анализ алгоритмов кластеризации в графах, а также описание и обоснование оригинальной двухэтапной модификации алгоритма Label Propagation (LPA), адаптированной под задачу устойчивого объединения семантически идентичных объектов.

Глава открывается формализацией понятия структуры объектов как графа, отражающего связи между потенциально идентичными объектами. Структура описывается в виде взвешенного графа $G = (V, E, w)$, где каждая вершина соответствует объекту, а рёбра отражают значения функции семантической

близости $F(i,j) \in [0,1]$, полученной, например, с помощью модели бинарной классификации.

Структурная интероперабельность, обеспечиваемая корректной идентификацией одинаковых объектов, позволяет выполнять агрегирование, сопоставление и консолидацию данных в динамично изменяющихся информационных системах. Выделяется идеальная (консонансная) структура — в которой рёбра соединяют только идентичные объекты, и реальная (ассонансная), содержащая ошибки двух типов: ложноположительные и ложноотрицательные связи. Задача алгоритма группировки — максимально приблизить наблюдаемую структуру к идеальной.

Также подчёркивается, что свойства устойчивости к ошибкам, масштабируемости и способности к инкрементальной обработке не принадлежат структуре как таковой, а являются характеристиками алгоритма, который эту структуру обрабатывает. Граф выступает в качестве универсальной модели, позволяющей работать с частичными и вероятностными связями, характерными для семантических идентичностей.

Далее проводится обзор и сопоставление наиболее известных алгоритмов кластеризации на графах, применяемых в задачах *community detection* и *entity resolution*. Сравнение осуществляется с учётом специфики задачи: зашумлённость графа, отсутствие знания числа кластеров, необходимость масштабируемости и адаптивности к динамике данных.

Анализируются следующие подходы: *edge betweenness* (Girvan–Newman), *modularity-based методы* (Louvain), *Label Propagation Algorithm* (LPA), а также Infomap, Walktrap и методы на основе спектральных или матричных разложений. Показано, что алгоритм LPA обладает наиболее подходящими характеристиками: он прост в реализации, хорошо масштабируется и не требует априорной информации о числе кластеров. Однако в базовой форме он подвержен проблеме залипания в локальные оптимумы и потере полноты в условиях зашумлённых связей.

Затем предлагается оригинальная двухэтапная модификация алгоритма LPA, направленная на устранение недостатков базового подхода. Суть модификации заключается в раздельной обработке сильных и слабых связей:

1. Первый этап — построение кластеров на подграфе, содержащем только рёбра с высокой степенью уверенности (веса $\geq \theta_{\text{strong}}$). Это

позволяет получить высокоточную начальную кластеризацию, устойчивую к шуму.

2. Второй этап — дообъединение неучтённых объектов на основе слабых связей (веса $\in [\theta_{\text{weak}}, \theta_{\text{strong}})$). Этот шаг направлен на повышение полноты кластеризации, позволяя присоединять слабо связанные объекты к уже сформированным группам, если общая вероятность связи с кластером превышает заданный порог.

Параметры θ_{strong} и θ_{weak} выбираются по результатам калибровки на размеченных данных: для сильных связей контролируется доля ложноположительных рёбер (не более 5%), а для слабых — доля ложноотрицательных (не более 20%).

Показано, что модифицированный алгоритм:

– сохраняет локальность и масштабируемость, что делает его пригодным для распределённых реализаций;
– обеспечивает адаптивность: новые объекты можно обрабатывать без полной переработки структуры;
– обладает устойчивостью к ошибкам, поскольку слабые связи используются только на втором этапе, после формирования устойчивых кластеров.

Таким образом, предложенный алгоритм сочетает в себе точность (за счёт строгой фильтрации связей на первом этапе) и полноту (за счёт мягкой агрегации на втором). Это делает его особенно эффективным для задач с неполными и вероятностными связями между объектами — таких, как группировка товарных позиций, документов или пользовательских идентификаторов.

Пятая глава диссертации посвящена экспериментальному подтверждению применимости предложенного метода динамичного формирования групп объектов по принципу идентичности. В главе описывается методика оценки качества, приводятся сравнительные результаты по нескольким алгоритмам, а также представлены примеры использования метода в реальных прикладных задачах.

В разделе 5.1 изложена методика оценки качества кластеризации. Для оценки эффективности алгоритмов был собран специальный тестовый набор данных из товарных предложений разных продавцов из одной категории. Набор состоит из около 3 тысяч объектов, вручную сгруппированных в 704 группы

(рис. 8). Это позволило использовать как информационно-энтропийные меры качества (однородность, полнота, V-мера), так и комбинаторные оценки качества (индекс Рэнда с поправкой на случайность, индекс Фаулкса–Мэллоуза). Такой подход дал возможность объективно сравнить качество группировки, независимо от количества кластеров и степени их пересечения.

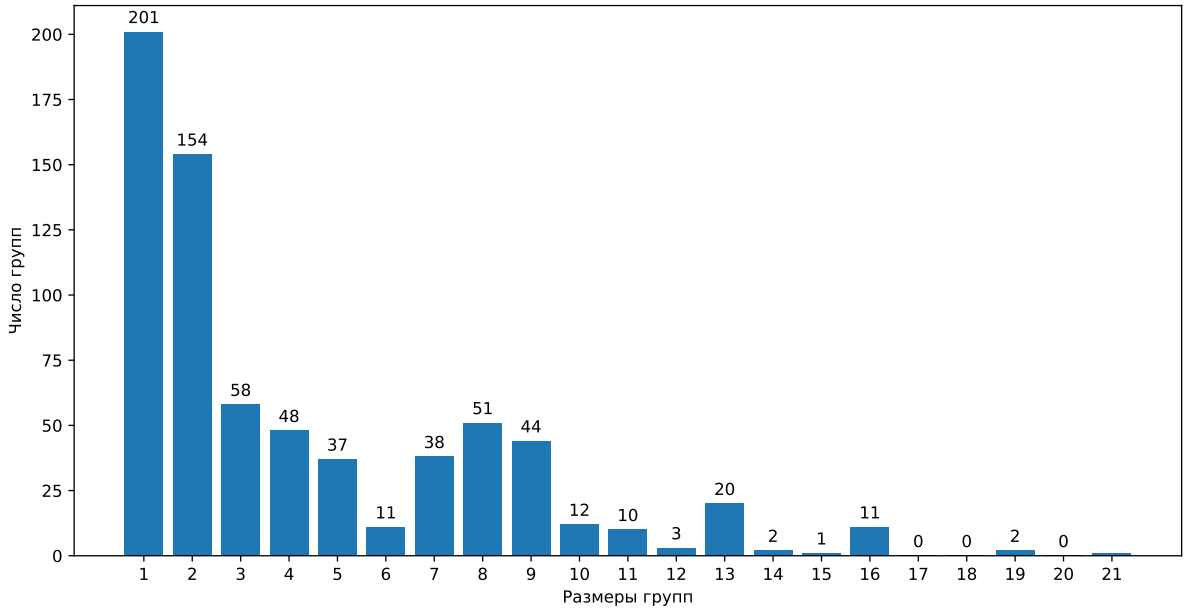

Рис. 8 — Распределение размеров групп в эталонной разметке.

Раздел 5.2 содержит численные результаты. Сравнивались различные подходы: транзитивное замыкание, алгоритм распространения меток (LPA), его модифицированный двухэтапный вариант, Louvain и DIANA. Отдельно рассматривались случаи работы на графах, полученных с различным порогом точности (95% и 80%), что позволило оценить влияние зашумлённости данных на итоговое качество. Результаты представлены в таблице 2.

Показано, что одностадийный LPA чувствителен к ошибочным связям: при увеличении числа слабых рёбер (граф 80%) он теряет однородность, но выигрывает в полноте. Напротив, при высокой точности графа (95%) он демонстрирует весьма высокую однородность (99.2%) при приемлемой полноте (91.6%). Двухэтапная модификация алгоритма (2-stage LPA) продемонстрировала наилучший баланс мер качества: высокая полнота (94.5%) при сохранении почти той же однородности (98.6%), что делает его наиболее эффективным в условиях ограничения по качеству исходных связей. DIANA показал высокую однородность ($\geq$ 99.1%) за счёт консервативной дивизивной стратегии, но

Метод	Однородность	Полнота	V-мера	Оценка Фаулкса-Мэллоуза	Индекс Рэнда с поправкой на случайность
Транзитивный подход (95%)	0.685	0.942	0.793	0.720	0.696
LPA (95%)	0.992	0.916	0.952	0.872	0.863
LPA (80%)	0.744	0.973	0.843	0.755	0.731
2-стадийный LPA	0.986	0.945	**0.965**	**0.882**	**0.871**
Louvain (95%)	0.993	0.902	0.945	0.868	0.855
Louvain (80%)	0.740	**0.975**	0.841	0.753	0.728
DIANA_1	**0.995**	0.879	0.933	0.840	0.819
DIANA_2	0.991	0.810	0.891	0.774	0.756

Таблица 2 — Результаты сравнения методов объединения объектов в группы

проиграл в полноте ($\leq 87.9\%$). Louvain, действующий на тех же графах, что и LPA, показал схожую с ним однородность, но меньшую полноту, что объясняется особенностями оптимизации модулярности и более осторожной стратегией включения объектов в кластеры.

В разделе 5.3 рассматриваются проблемы оценки качества работы системы группировки идентичных объектов в условиях больших и динамично обновляющихся данных. Показано, что стандартные меры качества, такие как полнота и однородность, становятся трудно применимыми в производственных условиях из-за невозможности получения полной эталонной разметки и высокой стоимости экспертной верификации.

Предложена практическая альтернатива в виде приближённой оценки однородности, основанной на двухэтапном семплировании пар объектов внутри групп, с балансировкой вклада малых и крупных групп в итоговую выборку. Такая методика позволяет выявлять отклонения от ожидаемого качества без необходимости полной проверки всех групп.

Для оценки полноты введено понятие покрытия (coverage). Это доля объектов, включённых в группы, содержащие как минимум два элемента. Данная мера качества пригодна для автоматизированного контроля и бизнес-интерпретации, так как наличие хотя бы одного идентичного объекта открывает возможности для объединения информации и улучшения пользовательского опыта. Также описана методика оценки максимально достижимого покрытия через экспертную проверку случайной выборки исключённых объектов, что позволяет адекватно интерпретировать значение меры качества в условиях отсутствия эталона.

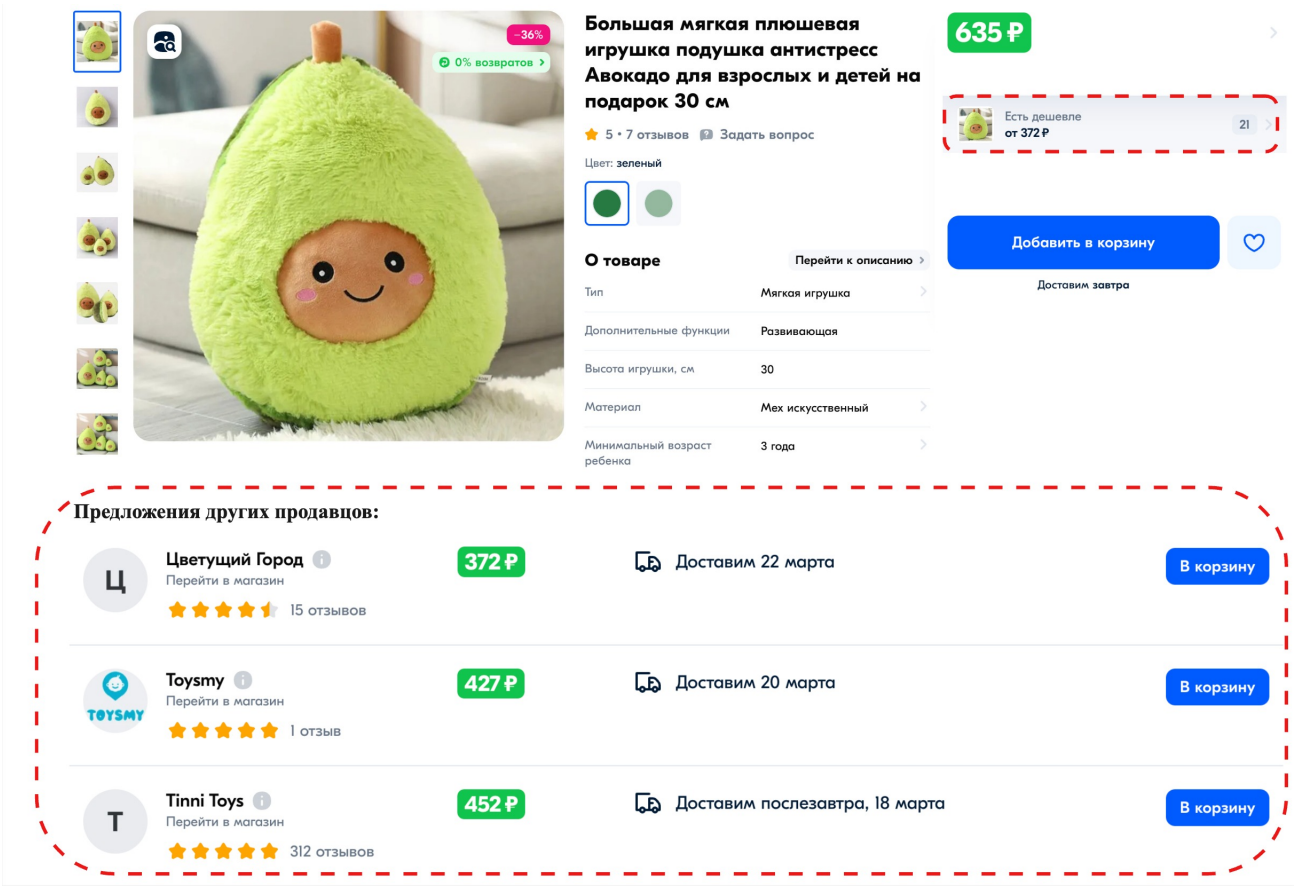

Рис. 9 — Блок с предложениями других продавцов и подсказка ”Есть дешевле”.

Раздел 5.4 иллюстрирует применение разработанного подхода на практике. Приведены примеры работы системы группировки товаров на платформе электронной коммерции, включая: формирование объединённых карточек товаров с предложениями от разных продавцов (рис. 9), подсказки о наличии аналогичного товара по более выгодной цене или с более быстрой доставкой, а также восстановление связей в пользовательском интерфейсе при исчезновении из продажи ранее отмеченного товара. Эти примеры подчёркивают бизнес-ценность метода: он не только обеспечивает консолидацию информации, но и напрямую влияет на пользовательский опыт и ключевые показатели маркетплейса.

Таким образом, пятая глава завершает диссертационное исследование, объединяя теоретические и инженерные аспекты. Она демонстрирует, что предложенный метод является не только теоретически обоснованным и вычислительно эффективным, но и демонстрирует его практическую реализуемость и прикладную значимость.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Изучены подходы к оптимизации аналитических SQL-запросов с использованием методов машинного обучения, который включает модели предсказания кардинальности и нейросетевые аппроксиматоры функции стоимости выполнения. Показано, что при условии значительных доработок предложенных в литературе методов, с помощью некоторых из них можно достичь ускорения выполнения сложных запросов в лабораторных условиях. Однако на практике в условиях динамично изменяющихся данных возникают проблемы устойчивости моделей к изменениям, а также высокие накладные расходы на их переобучение в условиях реального времени.
2. Разработан метод идентификации семантически идентичных объектов, основанный на комбинации бинарной классификации пар объектов и алгоритма кластеризации на графах. Произведён теоретический анализ нескольких алгоритмов выделения сообществ, а также сравнение качества алгоритмов по ключевым мерам качества (однородность и полнота) на реальных данных. Алгоритм распространения меток (LPA) выбран в качестве наиболее подходящего по ряду характеристик, а также по качеству работы. Предложенная архитектура адаптирована для распределённых систем с использованием парадигмы MapReduce, что обеспечивает масштабируемость и практическую применимость подхода в широком спектре приложений.
3. Разработан модифицированный двухэтапный алгоритм кластеризации на основе LPA, в котором реализована калибровка порогов идентичности с учётом допустимого уровня ошибок (5% и 20%). Показано, что данный алгоритм позволяет повысить полноту кластеризации без потери однородности, обеспечивая структурную согласованность групп. Предложенный метод апробирован на задаче поиска идентичных товаров на маркетплейсе. Предоставлены результаты работы системы.
4. Предложена и апробирована методика оценки качества группировки объектов, включающая показатели однородности и полноты, а также дополнительный показатель покрытия как альтернативу полноте в условиях отсутствия эталонной разметки. Эта методика адаптирована для анализа результатов динамической кластеризации в системах

обработки больших данных и продемонстрировала свою пригодность при проведении экспериментальных исследований.

Публикации в журналах из списка ВАК:

1. *Дулин С.К., Рябцев А.Б.* Анализ подходов к оптимизации запросов в аналитических СУБД // Образовательные ресурсы и технологии. №3 (44)’ 2023. С.73-80.
2. *Дулин С.К., Рябцев А.Б.* Алгоритм улучшения согласованности структурной интероперабельности // Надёжность. Том 24, № 2 (2024). С.8-16.
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

Прочие публикации:

1. *Дулин С.К., Рябцев А.Б.* Оценка планов выполнения SQL запросов для решения транспортных задач // Сетевой научно-методический журнал «Наука и технологии железных дорог», АО «НИИАС». 2023. Т. 7. №1 (25). С. 38-43.
2. *Дулин С.К., Рябцев А.Б.* Анализ концептов для проектирования базы геоданных железнодорожных геоописаний // Сетевой научно-методический журнал «Наука и технологии железных дорог», АО «НИИАС». 2024. Т. 8. №3 (31). С. 24-32.
3. *Рябцев А.Б., Дулин С.К.* Интеллектуализация анализа выполнения запросов в колоночной СУБД // Тезисы докладов 14-й Международной конференции «Интеллектуализация обработки информации». 2022. С. 103–105.
4. *Рябцев А.Б., Дулин С.К., Антипов И.Ф.* Подход к повышению согласованности структурной интероперабельности // Тезисы докладов 66-ой Всероссийской научной конференция МФТИ. 2024. С. 244–247.
5. *Рябцев А.Б., Дулин С.К.* Повышение структурной согласованности в задаче поиска групп идентичных объектов // Тезисы докладов 15-й Международной конференции «Интеллектуализация обработки информации». 2024. С. 33–35.