

ОТЗЫВ

Официального оппонента д.т.н. доцента Б. Г. Миркина

на диссертационную работу **Рябцева Антона Борисовича** на тему

**"Развитие метода динамического
формирования групп объектов
по принципу идентичности
для больших данных"**

представленную на соискание учёной степени
кандидата технических наук по специальности
2.3.8 - Информатика и информационные процессы

Актуальность темы обусловлена растущей потребностью в методах анализа и интеграции данных, способных работать в условиях их постоянной изменчивости. Классические алгоритмы кластеризации демонстрируют ограниченную применимость для задач динамического объединения семантически идентичных объектов: нужны новые подходы, учитывающие как высокую скорость поступления данных, так и необходимость структурной согласованности групп. Для современных информационных систем необходимо вычислительно эффективное и надёжное техническое решение данной задачи. Адаптация и развитие методов кластеризации на графах в составе комплексной методики оценки качества получаемой группировки является именно таким решением.

Научная новизна. Оригинальной и новой в работе является разработка метода динамического формирования групп объектов по «принципу идентичности» для больших данных, основанного на комбинации бинарной классификации пар объектов и модифицированного алгоритма адаптивного распространения меток (LPA), который, как показано, обеспечивает высокую структурную согласованность групп в условиях их постоянной изменчивости. Оригинальной является предложенная двухэтапная модификация алгоритма LPA, в которой раздельная обработка связей с высокой и средней степенью надёжности позволяет повысить полноту кластеризации без снижения её однородности. В работе описана новая методика оценки качества группировки для больших динамических данных, включающая подход к приближённой оценке однородности на основе двухэтапной выборки и показатель «покрытия» как альтернативу показателя «полноты» в условиях отсутствия эталонной разметки.

Достоверность и обоснованность результатов подтверждена математической корректностью разработанных методов, а также экспериментальной проверкой работы предлагаемых методов на тестовой выборке данных, успешным практическим внедрением результатов исследования в проектах отдела «Машинное обучение и матчинг» ООО «Озон

Технологии» и использованием результатов данной работы в рамках государственного задания номер 103-00001-25-02.

Основные положения диссертации обсуждались на международных и российских научных конференциях. Содержание диссертационной работы изложено в 3 работах, опубликованных в изданиях ВАК, одна из которых индексируется в Scopus и Web of Science, а также публикациями в других рецензируемых научных журналах.

Теоретическая значимость работы состоит в развитии теоретических основ и алгоритмического аппарата для задач объединения в группы семантически идентичных объектов в условиях больших и динамично изменяющихся данных. Теоретически значимыми результатами являются разработанный метод формирования групп объектов, который вносит вклад в теорию кластеризации за счёт обоснованного комбинирования бинарной классификации и теоретико-графовых построений. Предложенная двухэтапная модификация алгоритма распространения меток развивает теоретические представления о построении структурно согласованных группировок в графах с возможно ошибочными рёбрами. Разработанная методика оценки качества группировки расширяет теоретический инструментарий для анализа результатов кластер-анализа, предлагая новые подходы к проверке в условиях отсутствия полной эталонной разметки.

Практическая ценность: отдельные методы, представленные в работе, могут иметь широкое применение в различных отраслях для задач интеграции, очистки и консолидации больших динамичных данных. Результаты работы могут быть использованы: в сфере электронной коммерции, для автоматизированного управления ассортиментом и формирования объединённых карточек товаров на маркетплейсах, в логистике – для мониторинга и сопоставления состояния объектов в режиме реального времени, в задачах управления цифровыми двойниками сложных систем, в здравоохранении – для интеграции и очистки записей из разнородных медицинских регистров, а также в финансовом секторе – для повышения качества данных и дедубликации информации. Результаты работы внедрены и используются в системах ООО «Озон Технологии».

Диссертационная работа состоит из введения, пяти глав, заключения и списка литературы.

Введение содержит обоснование актуальности темы исследования, научной новизны и практической значимости результатов, полученных в работе, выносимые на защиту научные положения, обзор актуальных исследований по теме работы.

Первая глава посвящена теоретическим основам задачи динамического формирования групп объектов. В ней раскрываются ключевые понятия интероперабельности (способности систем к согласованному

взаимодействию) и структурной согласованности. Проводится обзор мер для оценки качества группировки и даётся строгая математическая постановка задачи, обосновывающая выбор комбинации бинарной классификации и графового алгоритма кластеризации.

Вторая глава фокусируется на задаче оценки схожести объектов — фундаментальном этапе, предшествующем группировке. Формализуется понятие функции схожести, анализируется влияние выбора порога схожести на структуру графа связей и рассматриваются типы ошибок (ложноположительные и ложноотрицательные). Особое внимание уделяется практической проблеме формирования признаков объектов, требующего в некоторых задачах выполнения ресурсоёмких аналитических SQL-запросов, что создаёт узкое место для производительности системы.

Третья глава представляет собой исследование проблемы медленного выполнения аналитических SQL-запросов. Анализируются традиционные методы оптимизации запросов и перспективные подходы на основе машинного обучения, включающие нейросетевые оптимизаторы. Делается вывод о практических ограничениях этих методов в условиях динамично меняющихся данных, обусловленных высокими затратами на поддержание актуальности моделей.

Четвёртая глава содержит описание и обоснование предлагаемого метода работы — модифицированного двухэтапного алгоритма кластеризации на основе алгоритма распространения меток (LPA). Предлагается подход, при котором на первом этапе формируются устойчивые ядра кластеров на основе сильных связей, а на втором этапе к ним присоединяются объекты по слабым связям для повышения полноты. Это позволяет получить более высокую полноту при том же уровне однородности кластеров, что и у одноэтапного варианта.

Пятая глава посвящена экспериментальной проверке предложенного метода. Описывается методика оценки качества на размеченных данных товарных предложений на маркетплейсе. Приводятся сравнительные результаты, демонстрирующие, что двухэтапный LPA превосходит другие алгоритмы по балансу однородности и полноты. Также предлагаются практические методы оценки качества в промышленных условиях, где полная разметка недоступна, и иллюстрируется прикладная ценность метода на примере онлайн-платформы для покупки товаров.

В заключении представлены основные результаты работы.

Замечания:

1. В главе 4 обоснование выбора алгоритма LPA перед другими методами является недостаточно строгим с математической точки зрения. В работе приведено сравнение по общим критериям (масштабируемость, отсутствие требования к числу кластеров),

однако отсутствует анализ устойчивости решений LPA к различным начальным условиям, не проведено теоретическое сравнение целевых функционалов (модулярность в Louvain и эвристическое правило большинства в LPA) применительно к рассмотренной задаче.

2. Критерии калибровки порогов τ_{strong} (5%) и τ_{weak} (20%) в разделе 4.3, хотя и продемонстрировали практическую эффективность, выбраны эмпирически. Не приведено анализа чувствительности качества группировки (V-меры) к вариациям этих порогов в окрестности выбранных значений. Целесообразно было бы представить график зависимости метрик от порогов, чтобы обосновать оптимальность именно этой пары параметров.
3. В главе 5 при сравнении алгоритмов не в полной мере раскрыта природа ухудшения однородности в LPA (80%). Утверждается, что это связано с накоплением ошибок, но отсутствует анализ матрицы смежности полученных кластеров: не показано, как именно растёт количество межкластерных рёбер при снижении порога и как это влияет на процесс распространения меток на структурном уровне.
4. Математический аппарат, связанный с консонансными и ассонансными структурами (Раздел 4.1), введён, но его связь с практическим алгоритмом LPA показана слабо. Не до конца ясно, как именно предложенный двухэтапный алгоритм минимизирует диссонанс в структуре графа и насколько формальная модель повлияла на проектирование алгоритма, кроме общей постановки задачи.

Следует отметить, что указанные замечания, в целом, не снижают значимости и научной ценности работы, не затрагивают её основных выводов, а являются скорее пожеланиями для дальнейшей работы.

Общее заключение: Диссертация Рябцева А.Б. является законченным научно-квалификационным исследованием, содержащим решение актуальных задач, объединённых общим подходом. Приведённые замечания не снижают общую положительную оценку выполненной автором работы. Структура и содержание диссертации соответствуют целям и задачам исследования. Основные результаты изложены в 8 печатных работах, из них 3 в журналах из списка ВАК.

Диссертационная работа Рябцева А.Б. «Развитие метода динамического формирования групп объектов по принципу идентичности для больших данных» соответствует пп. 9-14 «Положения о присуждении учёных степеней», утверждённого постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842, предъявляемым к диссертациям на соискание учёной степени кандидата технических наук по специальности

2.3.8 «Информатика и информационные процессы», а её автор, Рябцев А.Б., заслуживает присуждения ему учёной степени кандидата технических наук по данной специальности.

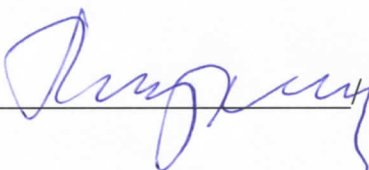
Официальный оппонент

доктор технических наук,

профессор департамента анализа данных и искусственного интеллекта
Федерального государственного автономного образовательного учреждения
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

10.11.2025 г.

 Б.Г. Миркин

