МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение высшего образования «Кемеровский государственный университет» (КемГУ)

650000, Кемерово, ул. Красная, 6
Телефон: 8(3842) 58-12-26. Факс: 8(3842) 58-38-85
E-mail: rector@kemsu.ru. http://www.kemsu.ru

18.02.2025 № 43/01.06

УТВЕРЖДАЮ
Проректор по
научно-исследовательской работе
ФГБОУ ВО «Кемеровский
государственный университет»,
д-р экономических наук,

___ _ Е.А. Жидкова

Отзыв ведущей организации

федерального государственного бюджетного образовательного учреждения высшего образования «Кемеровский государственный университет» на диссертационную работу Ахматшина Фарида Галиулловича на тему «Модели и алгоритмы автоматической группировки объектов для систем анализа и хранения данных на основе методов семейства *k*-средних», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1 – Системный анализ, управление и обработка информации, статистика

Актуальность темы исследования.

Работа посвящена повышению эффективности (вычислительной производительности, а также повышения качества результата по внешним и внутренним критериям качества) алгоритмов автоматической группировки объектов для систем анализа и хранения данных. Первая задача – разработке подхода к нормализации данных для предобработки входных данных, используемых в системах анализа данных. Вторая задача – разработке алгоритма кластеризации для системы анализа данных электрорадиоизделий. Третья задача – разработке алгоритма кластеризации для создания индекса векторной базы данных. Четвертая задача – разработке алгоритма автоматической группировки повторяющихся фрагментов блоков данных в системах хранения данных. Пятая задача – разработке процедуры инициализации центров кластеров для алгоритмов кластеризации при большом объеме данных. При этом увеличение объема обрабатываемых данных выявляет низкую вычислительную эффективность существующих алгоритмов. Таким образом, для решения поставленных задач требуется разработка новых алгоритмов автоматической группировки, используемых в системах анализа и хранения данных, на повышение эффективности алгоритмов кластеризации при обработке больших данных в системах автоматической группировки объектов, в том

числе в составе векторной СУБД и подсистем компрессии данных в составе систем хранения данных.

Общая характеристика работы

Диссертационная работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» и изложена 131 страницах, включая приложение. Основной текст состоит из введения, пяти разделов и заключения.

Во введении обоснована актуальность, поставлена цель и указаны задачи исследования, научная новизна и практическая значимость работы, изложены методы исследования, сформулированы основные положения, выносимые на защиту.

Раздел 1 посвящен вопросам нормализации данных в задаче автоматической группировки промышленной продукции на примере электрорадиоизделий (ЭРИ) по однородным производственным партиям, основанной на модели *k*-средних, а также разработке нового подхода по нормализации данных промышленной продукции, комбинирующего нормализацию по допустимым значениям параметров оцениваемых характеристик продукции и оценку Джеймса-Штейна.

Раздел 2 посвящен разработке алгоритма кластеризации электрорадиоизделий (решается задача разделения смешанной партии ЭРИ на однородные партии продукции) по результатам неразрушающих тестов с жадной эвристической процедурой выбора радиуса локальных концентраций по размеченным данным. Нами рассматривается задача поиска радиуса локальных концентраций (сгущений) в алгоритме кластеризации с заранее заданным числом кластеров. Результаты сравнивались с различными способами нормализации данных для решения задачи кластеризации методом *k*-средних.

Раздел 3 посвящен разработке нового алгоритма кластеризации, основанного на жадной агломеративной процедуре для построения индекса для векторной базы данных. В вычислительных экспериментах показано, что индекс (набор центров), сгенерированный этим алгоритмом, достигает более высокой оценки полноты по сравнению с алгоритмами *k*-средних и *k*-means++.

Раздел 4 посвящен разработке нового алгоритма автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма *k*-средних совместно с локально-чувствительным хэшированием (LSH) для использования в системах хранения данных.

Раздел 5 посвящен разработке новой процедуры инициализации центров кластеров для алгоритмов кластеризации, представляющей собой модификацию алгоритма инициализации *k*-means++. Полученная процедура инициализации применяется к большим данным. Учитывая меньшие вычислительные затраты, связанные с алгоритмом *k*-средних по сравнению с агломеративным алгоритмом, их применение в сочетании с модифицированным алгоритмом инициализации *k*-means++ к большим данным позволит уменьшить вычислительную сложность агломеративного алгоритма и улучшить его производительность.

В заключении сформулированы основные выводы и результаты, показано, что решение поставленных задач привело к достижению целей диссертации.

Достоверность и обоснованность научных положений, выводов и рекомендаций

В диссертационной работе Ахматшина Ф.Г. произведены исследования задачи автоматической группировки объектов для систем анализа и хранения данных, подробно исследованы алгоритмы семейства *k*-средних для решения данных задач.

Основные положения и результаты диссертационной работы докладывались и обсуждались на международных семинарах и конференциях: «Решетневские чтения» (2020 г., г. Красноярск), Mathematical Optimization Theory and Operations Research (MOTOR, 2023 – 2024 гг., г. Екатеринбург, г. Омск), семинар «Математические модели принятия решений» Институт математики имени С.Л.Соболева, (2024 г., г.Новосибирск).

Значимость результатов для науки

Теоретическая значимость состоит в дополнении эффективных алгоритмов решения задач автоматической группировки, а также алгоритмов предобработки данных для таких задач.

Практическая значимость полученных результатов

Как показывают приведенные в диссертационной работе исследования, разработанные подходы, алгоритмы и процедуры позволяют дополнить модельно-алгоритмический инструментарий, используемый в системах анализа данных результатов тестирования образцов промышленной продукции с повышенными требованиями качества, в частности – электронной компонентной базы космического применения, и могут использоваться в соответствующих испытательных технических центрах. Кроме того, новая процедура инициализации центров кластеров для алгоритмов кластеризации, имеет универсальный характер и может применяться при обработке больших данных в любых системах автоматической группировки объектов. Новые алгоритмы кластеризации применяются для построения индекса для векторной базы данных и для разработки модели оптимального использования дискового пространства с учетом компрессии данных.

Замечания к диссертационной работе

1. Предложенная новая процедура инициализации центров кластеров демонстрируют значительные преимущества по сравнению с известными алгоритмами инициализации. В качестве недостатка автореферата следует отметить, что в автореферате не приведены сравнительные результаты экспериментов с алгоритмом FAST k-means++ и k-means||.

2. Предварительная обработка данных, например, нормализация показателей, не является единственным решением для повышения качества решений последующих алгоритмов. Современные подходы комбинируют использование различных метрик и методов вычисления для поиска оптимальных решений, которые подходят для широкого спектра практических задач.

3. В тексте диссертации в таблицах 1.6 – 1.9 непонятно, какие данные из столбцов значений индекса Рэнда и целевой функции использованы для построения графиков на рисунках 1.1 – 1.8. - максимум, минимум или среднее.

4. В тексте диссертации:

– на стр. 81 непонятна причина изменения расстояния с 32КБ сразу на 64МБ при архивации повторяющихся блоков данных.

– на стр. 59 нет расшифровки аббревиатуры SIFT.

– на стр. 94 не ясно, почему в формуле (5.12) есть в левой части сумма по j от 1 до k.

5. В тексте автореферата также есть некоторое количество опечаток, например:

- на стр.14 в алгоритме 3 на Шаге 2 допущена опечатка: вместо S2 должно быть C2;

- на стр.16 опечатка в названии: «Алгоритм 4.2» заменить на «Алгоритм 4»;

- на стр.22 некорректно закончено предложение.

6. В автореферате на стр.12 таблица 2 перегружена данными. Непонятно, для чего приведены значения целевой функции: очевидно, что целевые функции в разных постановках задачи имеют различный смысл и не могут сравниваться напрямую.

Заключение о соответствии диссертации требованиям, установленным Положением о порядке присуждения ученых степеней

Несмотря на приведенные замечания, диссертационная работа Ахматшина Ф.Г. является завершенным научно-исследовательским трудом на актуальную тему, выполнена самостоятельно и на высоком научном уровне. Полученные автором результаты достоверны, выводы и заключения являются обоснованными.

Особую ценность работе придает её практическая направленность на решение практических задач систем анализа и хранения данных, на повышение эффективности алгоритмов кластеризации при обработке больших данных в системах автоматической группировки объектов, в том числе в составе векторной СУБД и подсистем компрессии данных в составе систем хранения данных.

Представленная диссертационная работа отвечает требованиям п.9 «Положения о порядке присуждения ученых степеней» Постановления правительства Российской Федерации от 24.09.2013 г. № 842, предъявляемым к кандидатским диссертациям, а её автор Ахматшин Фарид Галиуллович заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.1 – Системный анализ, управление и обработка информации, статистика.

Отзыв на диссертацию и автореферат обсужден и утвержден на заседании кафедры прикладной математики, протокол №_7__от «_17_» февраля_2025 г.

Заведующий кафедрой прикладной
математики, канд. техн. наук, доцент Каган Елена Сергеевна

Ведущая организация – ФГБОУ ВО «Кемеровский государственный университет».
Почтовый адрес: 650000, г. Кемерово, ул. Красная, 6.
Телефон: +7(3842)583845
E-mail: rector@kemsu.ru

«18» 02 2025 г.