АННОТАЦИЯ

диссертационной работы Касымовой Динары Тугелбековны на тему: «Исследование и разработка методов выявления и устранения противоречий в больших данных», представленной на соискание степени доктора философии (PhD) по специальности 6D070300 – Информационные системы

Актуальность темы исследования. В настоящее время обработка данных одна из самых актуальных проблем во многих областях науки и техники. Каждые два года объем оцифровки в мире удваивается. Объем информации в настоящее время составляет 33 зеттабайта, и ожидается, что к 2025 году объем данных в мире увеличится более чем в пять раз до 175 зеттабайт. Быстрый рост данных отражает потребность развития автоматических систем анализа и обработки данных.

В послании Президента страны Касым-Жомарта Токаева от 1 сентября 2020 года «Цифровизация - базовый элемент всех реформ»: Цифровизация - это не модный тренд, а главный инструмент достижения национальной конкурентоспособности. Работа с «данными» должна быть поднята на новый уровень. Одна из основных задач Правительства - обеспечение единой системы баз данных и их дальнейшее развитие. Мы приняли законы, которые позволяют Казахстану стать одним из международных центров обработки и хранения данных. Только в прошлом году в цифровой майнинг было инвестировано более 80 млрд тенге. Но мы не можем останавливаться на достигнутом, нам нужно привлекать цифровых гигантов со всего мира. В противном случае, по его словам, это сделают другие страны.

Выявление противоречий относится к проблеме поиска образцов данных, которые не соответствуют заданному понятию нормального поведения. Причем обнаружение новизны в данных, связанной с детектированием противоречий, обычно вначале включаются в нормальную модель после их обнаружения или отбрасываются. В своих трудах в различных областях зарубежные ученые их называют также аномалиями, выбросами, противоречивыми наблюдениями, исключениями: Лишуай Ли, Р. Джон Хансман (США), Сиянг Лу (США), Дж. Альбертенго, В. Хасан (Италия, Турин), Шигуанг Ван, Дексин Ю, Сяоганг Ма и Сюэ Син (Китай). В числе российских исследователей, внесших также вклад в теорию и практику анализа и устранения противоречий на латентно-семантической основе при выявлении противоречий в семантически близкой информации: Хомоненко А.Д., Агеев М.С., Дашонок В.Л., Добров Б.В., Краснов С.А., Кураленок, И.Е., Логашев С.В., Некрестьянов И.С., Лукашевич Н.В. и др.

Среди отечественных ученых, добившиеся научных результатов при обработке больших массивов данных в области распознавания изображений, обработки графических и лингвистических сигналов, классификации и кластеризации больших объемов данных различными методами: М.Н.

Калимолдаев, Е.Н. Амиргалиев, Р.И. Мухамедиев, Р.Р. Мусабаев, О.Ж. Мамырбаев и ряд других авторов.

Важность обнаружения в последующем противоречий обусловлена тем фактом, что они преобразуются в значительную информацию для использования в задачах управления различными процессами и объектами.

В последнее время большое внимание научной общественности уделяется исследованиям проблем транспорта в крупных городах, особенно, общественного транспорта, в том числе и в Казахстане, с использованием интеллектуального анализа и обработки больших данных.

Источниками больших данных являются, структурированные и неструктурированные данные, полученные с датчиков, видеокамер и установленных трекеров и т. д, информация о перемещении людей (с использованием сигналов GPS и Wi-Fi), базы данных зарегистрированных транспортных средств, расписание маршрутов общественного транспорта и т.д.

Основные задачи, решаемые с помощью интеллектуального анализа и обработки больших данных: прогнозирование дорожного движения и дорожно-транспортных средств с использованием современных методов и алгоритмов: статистических методов, k ближайших соседей (kNN), пространственно-временной корреляции, методов кластеризации и др. Вот некоторые авторы, которые активно занимаются транспортной проблематикой: Джеффри Дин (США), Санджай Гемават (Индия), А. Александров, Стефан Эвен, Макс Хеймел, Фабиан Хуэске (Германия), Одей Као, Фолькер Маркл, Эрик Найкамп, Даниэль Варнеке, Даниэль Келлехер, Р. Массобрио, С. Несмачнов (Уругвай), А. Черных, А. Аветисян, Г. Радченко (Россия).

Несмотря на большое количество статей и научных работ исследователей в этой области, из анализа текущего состояния исследований следует, что методы по обнаружению и устранению противоречий в больших данных для повышения эффективности и скорости работы систем недостаточно распространены. Учитывая также наличие острой потребности транспортных компаний Казахстана в исследовании и разработке методов выявления и устранения противоречий при обработке больших данных, было принято решение в качестве объекта исследования выбрать сферу общественного транспорта. Таким образом, тема диссертационной работы «Исследование и разработка методов выявления и устранения противоречий в больших данных» является актуальной.

Цель диссертационной работы. Исследование и разработка методов автоматического обнаружения и устранения противоречий в больших данных для повышения оперативности и эффективности принятия решений на основе статистической обработки и машинного обучения.

Задачи исследования:

1. Анализ существующих методов обнаружения и устранения противоречий в больших данных и выявление ключевых проблем.

2. Исследование и разработка комплексного метода автоматического обнаружения и устранения противоречий в больших данных, основанный на

статистических методах и алгоритмах машинного обучения, с учетом специфики предметной области.

3. Исследование и обоснование методики формирования эффективных статистических методов и алгоритмов машинного обучения для обнаружения и устранения противоречий с учетом специфики области исследования.

4. Создание интеллектуальной информационной системы для сферы общественного транспорта с использованием разработанного комплексного метода выявления и устранения противоречий в больших данных и исследование возможности обеспечения гибкого и быстрого изменения расписания автобусных маршрутов на ее основе.

Объект исследования: большие данные в сфере общественного транспорта.

Предмет исследования: методы и инструменты выявления и устранения противоречий в больших данных.

Методы исследования: методы математического и имитационного моделирования, методы кластеризации и классификации, статистические методы, машинное обучение, теория принятия решений, технологии разработки программного обеспечения.

На защиту выносится: Теоретический и сравнительный анализ обоснования комплексного метода автоматического обнаружения и устранения противоречий на основе статистических методов и алгоритмов машинного обучения, эффективность и точность которого отражается в результатах расчетов в рамках созданной интеллектуальной информационной системы общественного транспорта.

Научная новизна исследования.

1. Предложен комплексный метод автоматического выявления и устранения противоречий в больших данных на основе методов статистического анализа и машинного обучения и разработанной методики выбора его компонентов с учетом специфики предметной области.

2. Впервые были созданы архитектура, алгоритм и программа интеллектуальной информационной системы для выявления и устранения противоречий в больших данных городского общественного транспорта, повышающая оперативность и эффективность принятия решений при управлении общественным транспортом в городе.

3. Разработаны алгоритм и программа формирования шаблонов и управленческого решения для составления предварительного расписания городского автобусного маршрута в интеллектуальной информационной системе на основе очищенных больших данных общественного транспорта города.

Практическая значимость исследования: информационная система, основанная на аналитических и статистических исследованиях и алгоритмах машинного обучения, повышает эффективность общественного транспорта (автобусного парка) за счет рационализации интервалов маршрутов в результате автоматического обнаружения и устранения противоречий при формировании и обработке больших данных. Исследования также показали,

что можно повысить эффективность крупномасштабной обработки данных в области общественного транспорта.

На защиту выносятся научные положения: Теоретический и сравнительный анализ - это комплексный метод, основанный на автоматическом обнаружении и устранении противоречий на основе статистических методов и алгоритмов машинного обучения, эффективность и точность которого отражается в результатах расчетов в рамках созданной информационной системы общественного транспорта.

Личный вклад исследователя: Разработан комплексный метод выявления и устранения противоречий в больших данных. Проведены численные исследования и экспериментальная оценка предложенных моделей и алгоритмов. Разработана архитектура и программное обеспечение интеллектуальной информационной системы для сферы общественного транспорта с использованием разработанного комплексного метода для выявления и устранения противоречий данных.

Связь темы диссертации с планом научно-исследовательской работы: Диссертационная работа «Исследование и разработка инновационных информационных и телекоммуникационных технологий с использованием современных кибер-технических средств для интеллектуальной транспортной системы города» №AP05133699, руководитель проекта РГП главный научный сотрудник Института информационных и компьютерных технологий Комитета науки Министерства образования и науки Республики Казахстан, доктор технических наук, профессор Яворский В.В.

Структура и объем диссертации: Диссертация состоит из введения, трех глав, заключения и ссылок. Общий объем диссертации: 111 страницы письменного текста, в том числе 49 рисунок, 15 таблиц, 134 ссылки, 2 приложения.

Введение определяет актуальность работы и освещает ключевые вопросы, связанные с темой. Раскрыты идея работы, цели и задачи исследования, научная новизна и практическая ценность исследования, методы исследования.

В первом разделе дается широкомасштабное аналитическое исследование литературных и интернет-источников, показавшее тенденцию увеличения опубликованных научных работ в мире в связи с ростом глобальных данных, требующих развития теории и адекватных средств интеллектуального анализа для выявления конкретных идей и полезных моделей для принятия эффективных управленческих решений. Рассмотрены понятия противоречий в больших данных и типы противоречивых данных и проанализированы существующие методы и выявлены основные проблемы. Несмотря на то, что предварительная обработка данных является основным и важным этапом, показано, что проблемы выявления и устранения противоречий, часто возникают на более поздних этапах аналитики больших данных. Проведенный анализ существующих методов и инструментов выявления противоречий, их классификации, категорий и характеристик, показал наличие множества проблем при анализе и интеграции данных.

Приведены эффективные примеры использования алгоритмов машинного обучения и статистических методов для выявления противоречий. При этом основная проблема заключается в необходимости выбора методов и технологий соответствующих конкретным целям и задачам анализа и предметной области, в связи с чем для исследования была выбрана конкретная предметная область больших данных - городской пассажирский транспорт.

Во втором разделе обоснована и предложена функциональная схема аналитики больших данных с подсистемой комплексного метода выявления противоречий и их автоматического устранения. Особенность предложенной функциональной схемы заключается в том, что выбор используемого в каждой компоненте метода обосновывается особенностями и целью обеспечения заданных параметров идентификации построенных на их основе моделей для решаемой задачи предметной области - аналитики больших данных городского пассажирского транспорта. Обоснованные и выбранные для предметной области единые статистические методы используются на следующих этапах: получение и предварительная обработка данных, формирование базы больших данных, а также обнаружение и устранение противоречий в базах больших данных. Основные направления предварительной обработки данных сгруппированы по типам с учетом текущих проблем в каждом из них. На основе их анализа были определены методы, которые можно использовать для обработки данных, содержащих противоречия. Сюда входят методы удаления данных из исходных из заданного набора данных, поэтому они были проанализированы и изучены для использования для заданной предметной области. Проведен экспериментальный сравнительный анализ статистических критериев для выбора для создания эффективного алгоритма вывода законов изменения переменной общественного транспорта по отношению к предиктору. Критерий Граббса был выбран потому, что он выявляет противоречивые данные с высокой точностью и может работать с большими выборками, а также были детально рассмотрены алгоритмы k-means, алгоритмы SVM, DBSCAN и получены экспериментальные результаты. Согласно результатам, полученным путем сравнения алгоритмов кластеризации, обнаружение противоречивых данных с использованием алгоритма k-средних показало более высокий процент, а время вычисления было быстрее. Также осуществлен подбор компонентов машинного обучения многослойной нейронной сети - построение градиентных изображений и разработан соответствующий алгоритм обучения. Сравнивались и оценивались точность алгоритмов при различных значениях параметров противоречия. Результаты, полученные в разном процентном соотношении, подтверждают эффективность использования предложенного комплексного метода.

В третьем разделе разработана структура подсистемы комплексного метода обнаружения и устранения противоречий в больших данных в сфере общественного транспорта, компоненты которой обоснованы во втором разделе. Отличительной особенностью предложенного комплексного метода является реализация: 1) двухуровневой системы обнаружения противоречий: с использованием статистического критерия Граббса и проверкой алгоритма k-

средних. 2) обучения в нейронной сети для обнаружения противоречивых и удаления данных из информации, полученной в течение определенного периода времени. Впервые предложена архитектура интеллектуальной информационной системы общественного транспорта с подсистемой для выявления и устранения противоречий в больших данных. Отличительной особенностью этой архитектуры является наличие трехкомпонентной нереляционной базы данных MongoDB для хранения больших объемов данных, а также использование методов обработки, которые позволяют обрабатывать информацию распределенным образом на нескольких физических серверах. Сбор и предварительная обработка больших данных осуществляется в нереляционной базе данных MongoDB в режиме реального времени. Модуль для глубокого обучения нейронной сети предназначен для формирования шаблонов и управленческого решения в виде предварительного расписания городского автобусного маршрута для каждого периода времени. Были проведены численные эксперименты по фактическим данным по пяти автобусным маршрутам, результаты которых показали правильность выбора критерия Граббса для решения задач обработки данных для управления автобусным транспортом. Результаты исследований по выявлению противоречий в данных городских автобусных маршрутов показали эффективность созданной информационной технологии и перспективы ее дальнейшего использования и развития. По результатам было предложено внести изменения в расписание движения общественного транспорта. Проведена экспериментальная проверка точности получаемых моделей в процессе их обработки и анализа, показавшая отсутствие противоречивых значений очищенном наборе данных. Таким образом, было доказано, что набор данных, полученный после исключения противоречивых данных обладает гарантированной точностью, и может быть использован в сфере общественного транспорта.

В заключении описаны основные результаты и выводы диссертации.

Апробация работы. Результаты диссертации были представлены и обсуждены на следующих научно-методических конференциях и семинарах: Международная конференция IEEE по мягким вычислениям и измерениям (SCM 2017, 6 июля 2017 г., Санкт-Петербург; Российская Федерация), XLVII Международная научная конференция «Наука вчера, сегодня, завтра» - практическая конференция (Новосибирск, 2017), Международный Азиатский семинар «Система оптимизации сложных систем» (Кыргызстан, 2018, 2019), IV Международная научно-практическая конференция «Информатика и прикладная математика» (26-29 сентября, Алматы, 2018 г.), Научная конференция «Инновационные ИТ и умные технологии», посвященная 70-летию профессора И.Т. Утепбергенова (Алматы, 2019 г.), IV Международная научно-практическая конференция «Информатика и прикладная математика», посвященная 70-летию Т.Н. Биярова и В. Войцука, 60-летию Е. Н. Амиргалиева (Алматы, 25-29 сентября 2019г.), научные семинары Института информационных и компьютерных технологий МОН РК.

Публикация результатов. Основные научные результаты диссертации опубликованы в 19 публикациях, в том числе 8 - в научных изданиях, рекомендованных Комитетом по контролю в сфере образования и науки Министерства образования и науки Республики Казахстан, 1 - в научных публикациях, включенных в международную базу данных Scopus, 10 - в международных научных конференциях.

1. Tashev A., Kuandykova J., Kassymova D., Akhmediyarova A. Detection and elimination of discrepancies in big data at transport applying statistical methods // Journal of Theoretical and Applied Information Technology. – 2020. - Vol.98. №9. – P.1435-1445. (Scopus, процентиль - 37)

2. Утепбергенов И.Т., Касымова Д.Т., Ахмедиярова А.Т., Ескендирова Д.М. Подход к выявлению и устранению семантических противоречий в «больших данных» // Вестник КазАТК имени М.Тынышпаева. – Алматы, 2017. - №2 (101). – С.200-206.

3. Касымова Д.Т., Ескендирова Д.М., Ахмедиярова А.Т. Үлкен деректердегі қайшылықтарды жою және анықтау // Вестник КазАТК им. М.Тынышпаева. – Алматы, 2017. - № 3 (102). - С.76-80.

4. Utepbergenov I.T., Kassymova D.T., Musabekov N.R., Utegenova A.U., Muslimova A.K. Integrated Approach for Implementing the Virtual Information Infrastructure of the automated process control system // Вестник КазНУ им. Аль-Фараби. – Алматы, 2015. - № 3(86). – С. 152-156.

5. Утепбергенов И.Т., Ахмедиярова А.Т., Касымова Д.Т. О задаче моделирования регулярного города с помощью сети Петри // Вестник КазАТК им. М.Тынышпаева. Серия «Информационные системы». - Алматы, 2016. - №1 (96). – С.77-81.

6. Қасымова Д.Т., Ахмедиярова А.Т., Шаяхметова А.С., Тұрдалыұлы М. Үлкен деректерде кездесетін қайшылықтарды анықтау мен жоюға қолданылатын әдістерге талдау // ҚазҰТЗУ хабаршысы. «Техникалық ғылымдар» сериясы. – Алматы, 2020. - № 2 (138). – С. 487-495.

7. Қасымова Д.Т., Ахмедиярова А.Т., Бижанова А.С. О задаче управления светофорами на перекрестках // Вестник КазАТК им. М.Тынышпаева. Серия «Информационные системы». – Алматы, 2016. - №1(96). - С.74-77.

8. Хомоненко А.Д., Касымова Д.Т., Куандыкова Д.Р., Ахмедиярова А.Т. Проблемы устранения противоречий в больших данных // Вестник КазНИТУ. Серия «Технические науки». - Алматы, 2019. -№2(132). - С.418-424.

9. Касымова Д.Т., БуранбаеваА.И., Нуркаманова М.А. Сравнительный анализ современных систем моделирования городского транспорта // Вестник КазНИТУ. Серия «Технические науки». - Алматы, 2018. - №4(128). - С.170-176.

10. Khomonenko A.D, Dashonok V.L., Kassymova D.T., Ivanova K.A. Approach to processing of data from social networks for detecting public opinion on quality of educational services // Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017. - St. Petersburg, 2017. - № 7970707. - P. 737-739.

11. Khomonenko A.D., Khalil M.M., Kassymova D.T. Probabilistic models for evaluating the performance of cloud computing systems with web interface // SPIIRAS Proceedings, 2016. - Issue 6(49). - P.49-64.

12. Дашонок В.Л., Касымова Д.Т. Проблемы устранения противоречий в больших данных // Сборник статей по матер. XLVII междунар. науч.-практич. конф. «Наука вчера, сегодня, завтра». – Новосибирск: СибАК, 2017. -№ 6(40). – С. 22-26.

13. Утепбергенов И.Т, Ахмедиярова А.Т., Касымова Д.Т. Внедрение интеллектуальных систем на транспорте // Матер. V междунар. науч.-практич. конф. "Информатика и прикладная математика". — Алматы: ИИВТ МОН РК, 2020. - С. 332-340.

14. Ахмедиярова А.Т., Касымова Д.Т. Анализ изменчивости времени с использованием больших данных мегаполиса // Матер. III междунар. науч. конф. «Информатика и прикладная математика», посв. 80-летнему юбилею проф. Бияшева Р.Г. и 70-летию проф. Айдарханова М.Б. – Алматы, 2018. – Т. 1. – С. 207-212.

15. Ахмедиярова А.Т., Касымова Д.Т., Нуркаманова М.А. Использование генетического алгоритма в управлении светофором // Матер. XIV междунар. Азиатской школы-семинара «Проблемы оптимизации сложных систем». – Кыргызская Республика, 2018. – Т. 2. – С. 388-392.

16. Касымова Д.Т., Утепбергенов И.Т., Ескендирова Д.М. Современные подходы обработки больших объемов данных // Труды междунар. Сатпаевских чтений «Роль и место молодых ученых в реализации новой экономической политики Казахстана». - Алматы, 2015. - Том IV. - С. 208-212.

17. Касымова Д.Т., Утепбергенов И.Т., Ахмедиярова А.Т. Предварительная обработка больших данных: методы и перспективы // Матер. науч. конф. ИИВТ КН МОН РК «Инновационные IT и Smart-технологии», посв. 70-летнему юбилею профессора Утепбергенова И.Т. – Алматы, 2019. – С. 256-264.

18. Ахмедиярова А.Т., Касымова Д.Т. Обзор методов сбора данных для интеллектуальных транспортных систем // Матер. IV междунар. науч прак. конф. «Информатика и прикладная математика», посв. 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию проф. Амиргалиева Е.Н. – Алматы, 2019. - С. 425-432.

19. Калимолдаев М.Н., Яворский В.В., Сонькин М.А., Вуйцик В., Утепбергенов И.Т., Ахмедиярова А.Т., Касымова Д.Т., Клюева Е.Г., Байдикова Н.В., Есмагамбетова М.М. Формирование хранилища и анализ больших данных передвижений в городе // Сборник матер. VI междунар. науч.-практич. конф. «Big Data и анализ высокого уровня». — Минск: Бестпринт, 2020. - С. 59-70.

Свидетельство о внесении данных в государственный реестр прав на объекты авторского права:

1. А.с. 12585. «Методы выявления и устранения противоречий» / Д.Т. Касымова, А.А. Ташев, И.Т. Утепбергенов, А.Т. Ахмедиярова; опуб.15.10.2020, - 1с.