# КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

## COMPUTER SIMULATION HISTORY

УДК 004.04
DOI 10.18413/2687-0932-2020-47-4-792-802

# Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных

**Багутдинов Р.А., Саргсян Н.А., Краснoплахтыч М.А.**
Профессиональная образовательная организация частное учреждение
«Автомобильно-дорожный колледж»,
Россия, 354051, г. Сочи, ул. Яна Фабрициуса, д. 26а/1
E-mail: rav379@mail.ru

**Аннотация**
В данной работе представлен аналитический анализ и инструменты обработки больших данных, рассмотрены некоторые аспекты интеллектуального анализа различных данных с высокой интенсивностью, форматов и происхождения, неоднозначности, избыточной или недостаточной полноты значений. Авторами рассматриваются актуальные аспекты соответствующего системного и аналитического анализа, рассматривается потенциал слияния разнородных данных. Предоставлены инструменты и традиционные методы интеллектуального анализа данных. Обозначены проблемы разрыва данных, обнаружение выбросов, выявление аномалий данных, непрерывный аудит, стратегии вычислительных кластеров, их аспекты и описание. Выявлено, что проблемы больших данных включают не только разномасштабность данных, но и неоднородность, отсутствие структурной интеграции, качество данных, конфиденциальность, безопасность. Результаты данной работы могут быть полезны в научных изысканиях исследователей, которые сталкиваются с проблемами обработки больших разнородных и разномасштабных данных в различной сфере научных знаний и областей практического применения.

**Ключевые слова**: обработка больших данных, разнородные данные, разномасштабные данные, интеллектуальный анализ, аналитика данных, экспертные системы, системы обработки данных.

**Для цитирования:** Багутдинов Р.А., Саргсян Н.А., Краснoплахтыч М.А. 2020. Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных. Экономика. Информатика. 47 (4): 792–802. DOI 10.18413/2687-0932-2020-47-4-792-802.

# Analytics, tools and intellectual analysis of large different and differential data

**Bagutdinov R.A., Sargsan N.A., Krasnoplakhtych M.A.**
Professional educational organization private institution
"Automobile and road college", 26a/1 Yan Fabricius, St, Sochi, 354051, Russia
E-mail: rav379@mail.ru

**Abstract**
This paper presents analytical analysis and tools for processing big data, considers some aspects of mining various data with high intensity, formats and origin, ambiguity, excessive or insufficient completeness of values. The authors consider the relevant aspects of the relevant system and analytical analysis, and consider the potential for fusion of heterogeneous data. Tools and traditional data mining techniques are provided. The problems of data disruption, outlier detection, data anomaly detection, continuous audit, computational cluster strategies, their aspects and description are identified. It was revealed that the problems of big data include not only different scales of data, but also heterogeneity, lack of structural integration, data quality, confidentiality, security, etc. The results of this

work can be useful in scientific research of researchers who are faced with the problems of processing large heterogeneous and multi-scale data in various fields of scientific knowledge and areas of practical application.

**Keywords**: big data processing, heterogeneous data, multi-scale data, data mining, data analytics, expert systems, data processing systems.

**For citation:** Bagutdinov R.A., Sargsan N.A., Krasnoplakhtych M.A. 2020. Analytics, tools and intellectual analysis of large different and differential data. Economics. Information technologies. 47 (4): 792–802 (in Russian). DOI 10.18413/2687-0932-2020-47-4-792-802.

## Введение

Существуют разрывы между большими данными и текущими возможностями анализа данных в системах непрерывного аудита. В частности: огромный объем, высокая скорость и огромное разнообразие вводят пробелы (разрывы) в согласованности данных, идентификации данных и агрегировании данных для связи баз данных в системе непрерывного аудита. Каждый из пробелов создает соответствующие проблемы, указанные в таблице 1. Идентификация данных относится к записям, которые связывают два или более отдельно записанных фрагмента информации об одном и том же человеке или объекте [Багутдинов, 2014, Багутдинов, 2019]. Когда данные структурированы, идентификация проста. Однако идентификация становится затруднительной при аудите больших данных, где большая часть данных может быть неструктурированной. Отсутствие целостности данных обычно связано с недостоверными данными и неполными данными [Багутдинов, 2018, Багутдинов, 2019]. Согласованность данных является наиболее важной проблемой для систем непрерывного аудита больших данных и связана с взаимозависимыми данными между приложениями и всей организацией. Конфиденциальность данных означает, что определенные данные или связи между точками данных являются конфиденциальными и не могут быть переданы другим [Островский, 2018, Островский 2019]. В эпоху больших данных данные могут быть легко связаны с другими данными. После утечки некоторых конфиденциальных данных они могут распространяться с высокой скоростью и связываться с большим количеством связанных данных [Островский, 2020]. Таким образом, конфиденциальность больших данных становится еще более актуальной и важной для сохранения имиджа организации (использующей данные) и обеспечения конкурентного преимущества. Агрегация данных необходима для нормальной работы непрерывного аудита с использованием больших данных для значительного суммирования и упрощения больших данных, которые, поступают из разных источников [Kabacoff, 2015].

Таблица 1
Table 1

Аналитика больших данных
Big data analytics

| Особенности больших данных | Проблемы в непрерывном аудите | Большой разрыв данных |
|---|---|---|
| • объем<br>• разнообразие<br>• скорость<br>• достоверность | • конфликтующие данные,<br>• неполные данные,<br>• данные с различными идентификаторами,<br>• данные в разных форматах,<br>• недостоверные данные,<br>• асинхронные данные,<br>• поиск зашифрованных данных,<br>• аудит зашифрованных данных,<br>• аудит агрегированных данных | • идентификация,<br>• целостность,<br>• согласованность,<br>• конфиденциальность,<br>• агрегирование данных |

Неоднородность является одной из важных характеристик больших данных. Данные из разных и различных источников по своей природе обладают множеством различных типов и форм представления, и они могут быть взаимосвязаны, невзаимосвязаны, и представлены непоследовательно. Под обработкой неоднородных больших данных здесь будем также понимать обработку структурированных, полуструктурированных и даже полностью неструктурированных данных одновременно [Zhang, Yang, Appelbaum, 2015]. Большие данные связывают большие объемы и сложные наборы данных с несколькими независимыми источниками. Анализ больших данных может быть проблематичным, поскольку он часто включает сбор и хранение смешанных данных, основанных на различных закономерностях или правилах. Здесь большую роль имеет контекст данных, их описание. Например, существующие данные в производстве не имеют никакого отношения к контексту об истории, расписании, привычках, задачах и местоположении пользователей и т. д. В контексте больших данных контекстуализация может быть привлекательной парадигмой для объединения разнородных потоков данных для улучшения качества процесса добычи или классификации. Кроме того, контекстное описание данных, безусловно, сокращает время обработки и потребление ресурсов путем концентрации процессов генерации больших данных (например, мониторинга реальных ситуаций с помощью различных систем) только на источниках, которые, как ожидается, будут наиболее перспективными (взаимосвязанными) в зависимости от конкретного контекста. Тут стоит выделить три парадокса больших данных [Tak, Gumaste, Kahate, 2015]:

1. Парадокс идентичности – мы стремимся большие данные идентифицировать, выставить метки для простоты их дальнейшей обработки, но это также угрожает их идентичности, может вызвать дополнительные ошибки и неточности в конечном принятии решений. Это парадокс идентичности. Выделение одних данных в отдельную группу может привести к преувеличению групповых (кластерных) различий данных, применение различных методов и специфик обработки этих групп согласно их идентификации, но это же приводит к тому, что незначительные изменения внутри этих групп могут повлиять на конечное решение задачи и привести к множеству различных решений, возрастанию объема данных и усложнению самого процесса обработки разнородных данных.

2. Парадокс прозрачности – аналитика больших данных зависит от небольших данных. Небольшие входные данные агрегируются для получения больших наборов данных. Этот сбор данных происходит незаметно. Большие данные используют эти данные, чтобы сделать обработку более прозрачной, понятной; но в основном все процессы происходят по принципу «черного ящика»; и некоторые процессы, инструменты и методы «непрозрачны», скрыты слоями физической, юридической и технической конфиденциальности.

3. Парадокс власти – датчики больших данных и большие пулы данных находятся преимущественно в руках влиятельных посреднических учреждений, а не обычных людей. Конфиденциальность, автономность, прозрачность и защита личности с самого начала не встроены в большие данные.

Вообще говоря, значения, скрытые в больших данных, зависят от «свежести» данных. Следовательно, должен быть разработан принцип важности, связанный с аналитической ценностью, чтобы решить, какие данные следует отбрасывать, а какие хранить. Для решения проблемы анализа больших данных необходимо выполнить следующее:

1. Загрузка данных – должно быть разработано программное обеспечение для загрузки данных из нескольких и различных источников данных. Система должна иметь дело с распределенной природой данных с одной стороны и нераспределенной природой источника данных. Система должна иметь дело с поврежденными записями и должна предоставлять услуги непрерывного мониторинга.

2. Анализ данных – большинство источников данных предоставляют данные в определенном формате, который необходимо проанализировать. Некоторые форматы, такие как JSON, сложно анализировать, поскольку запись может содержать много строк текста, а не только одну строку на запись.

3. Аналитика данных – решение для анализа больших данных должно поддерживать быстрые итерации для правильного анализа данных.

Типы аналитических методов больших данных включают: описательную аналитику (включающую описание и обобщение моделей знаний); прогнозируемую аналитику (т. е. прогнозирование и статистическое моделирование для определения будущих возможностей и дальнейшей обработки); предписывающую аналитику (т. е. некий программный модуль, который позволит помочь аналитикам в принятии решений на основе данных путем определения действий и оценки их воздействия).

Для обработки больших данных используются распределенные системы, базы данных с массивной параллельной обработкой, нереляционные базы данных или базы данных в памяти. Базы данных параллельной обработки обеспечивают высокую производительность запросов и масштабируемость платформы. Нереляционные базы данных, такие как Not Only SQL, используются для хранения и управления неструктурированными или нереляционными данными и предназначены для масштабирования, гибкости модели данных и упрощенной разработки и развертывания приложений. Базы данных в памяти управляют данными в памяти сервера, обеспечивая возможность ответов в режиме реального времени из базы данных. Кроме того, базы данных в памяти используются для расширенной аналитики больших данных, особенно для ускорения доступа и анализа аналитических моделей [Schotma, Mitwalli, 2013; Jaseena, David, 2014; Kreuter and al., 2015]. Помимо прочего, есть несколько инструментов для работы с большими данными, таких как Hive, Splunk, Tableau, Talend, RapidMiner и MarkLogic. Hive упрощает управление и запрос больших массивов данных, находящихся в распределенном хранилище. Splunk фокусируется на использовании машинных данных, созданных из различных источников, таких как датчики и веб-сайты. Tableau – это инструмент визуализации данных, который позволяет пользователям создавать точечные диаграммы, графики и карты. Talend – это инструмент с открытым исходным кодом для разработки, тестирования и развертывания продуктов для управления данными и интеграции приложений. RapidMiner предоставляет предприятиям централизованное решение с мощным и надежным графическим пользовательским интерфейсом, который позволяет пользователям создавать, поддерживать и предоставлять прогнозную аналитику. MarkLogic может использоваться для обработки больших объемов данных и предоставления пользователям доступа к ним через обновления и оповещения в режиме реального времени [Chen, Mao, Liu, 2017].

Когда скорость становится очень высокой, инструменты с большими данными, вероятно, будут единственным вариантом. Инструменты больших данных способны очень быстро извлекать и анализировать данные из огромных наборов данных, что особенно полезно для быстро меняющихся данных, которые можно анализировать с помощью обработки в памяти. Инструменты больших данных способны распределять сложные задания обработки по большому количеству узлов, уменьшая сложность вычислений [Elgendy, Elragal, 2014]. Oozie и Elastic MapReduce (EMR) с Flume и Zookeeper используются для обработки объема и достоверности данных, которые являются стандартными инструментами управления большими данными [Yusuf, 2017]. MapReduce работает с числовыми и номинальными значениями, однако, алгоритмы должны быть переписаны, и требуется понимание системного проектирования. С помощью YARN Hadoop теперь поддерживаются различные модели программирования, а также почти в реальном времени и в пакетном режиме. Существует множество качественных программных инструментов, позволяющих воспользоваться преимуществами больших данных. Например, Kitenga Analytics Suite от Dell – ведущая в отрасли платформа для поиска и анализа больших данных, которая была разработана для интеграции информации всех типов в легко развертываемые визуализации. Этот инструмент позволяет интегрировать разнородные источники данных и экономически эффективно хранить растущие объемы данных. Kitenga может напрямую анализировать результаты обработки данных, используя инструменты визуализации информации, которые напрямую связаны с файлами, а также индексировать созданные данные и

метаданные в форму с возможностью поиска со встроенными возможностями визуализации [Schotman, 2013]. Часто при обработке больших данных используется одна из следующих трех стратегий [Jaseena, David, 2014]:

1. Внутренний вычислительный кластер. Для долговременного хранения уникальных или конфиденциальных данных часто имеет смысл создавать и поддерживать кластер Apache Hadoop, используя серию сетевых серверов во внутренней сети организации.

2. Внешний вычислительный кластер. В отрасли информационных технологий наблюдается тенденция к передаче элементов инфраструктуры сторонним поставщикам услуг. Некоторые организации упрощают для системных администраторов аренду готовых кластеров Apache и систем хранения данных.

3. Гибридный вычислительный кластер. Распространенным гибридным вариантом является предоставление ресурсов внешнего вычислительного кластера с использованием сервисов для задач анализа больших данных по требованию и создание внутреннего компьютерного кластера для долгосрочного хранения данных.

Анализ больших данных включает в себя несколько отдельных этапов и сопутствующих проблем, некоторые из которых показаны в таблице 2, некоторые проблемы выходят за рамки настоящей работы. Помимо общих технических проблем больших данных, существуют дополнительные проблемы: сделать данные более доступными путем структурирования и добавления метаданных с учетом интеграции отдельных хранилищ данных; решение нормативных вопросов, касающихся владения данными и конфиденциальности данных, в том числе, если мы имеем виду обработку закрытых данных, например, в военной отрасли.

Проблемы больших данных включают не только разномасштабность данных, но и неоднородность, отсутствие структурной интеграции, качество данных, конфиденциальность, безопасность и т. д. Для достижения качественных результатов необходимо использовать целостный, комплексный, системный подход к управлению данными, их анализу и информации.

Таблица 2
Table 2

Аспекты и описание аналитики больших данных
Big data analytics aspects and description

| Аспекты | Описание |
|---|---|
| • основные шаги в анализе больших данных | • получение, запись, очистка, извлечение, интеграция, агрегация, представление, анализ, моделирование, интерпретация |
| • проблемы во время выполнения шагов обработки | • неоднородность, различие во времени (или несостыковка временных интервалов при получении данных), масштаб, конфиденциальность |

Даже если корреляция может оказаться надежной в течение определенного периода времени, аналитика больших данных сама по себе не может дать представление о том, что может привести к нарушению корреляции, или о том, какая модель может появиться на ее месте. Критика аналитики больших данных заключается в том, что существование массивных наборов данных не устраняет традиционные статистические ошибки выборки и смещения выборки. Развитие и повсеместность сенсорных сетей и других источников больших данных вносит свои коррективы в существующие традиционные методы и способы обработки данных. Всё это оказывает влияние на следующее поколение технологий больших данных:

- Глобальный рост сети Интернет, большой объем хаотичных неструктурированных разнородных данных. По мере того, как все больше пользователей подключаются к сети, технологии больших данных должны будут обрабатывать большие объемы данных.

- Обработка в режиме реального времени. В последние годы стали доступны системы потоковой обработки, такие как Apache Storm, которые обеспечивают новые возможности приложений. Здесь стоит остановиться подробнее на актуальных проблемах. Требуется принятие быстрых решений в определенные моменты времени получения непрерывных потоков данных. Такая обработка накладывает специфические требования к методам обработки. Трудность заключается в выявлении конкретных точек отчета (периодов), которые должны иметь максимальное количество информации для получения соответствующего решения, при этом необходимо избегать избыточности данных. Также не совсем понятно какой минимальный и максимальный набор выборки данных нужно получить, чтобы выявить ту или иную закономерность в данных, аномалию или решить требуемую задачу.

- Обработка сложных типов данных. Сложность и суть данных никоим образом не должны влиять на скорость работы алгоритмов и методов больших данных. Вновь создаваемые методы должны легко обрабатывать такие данные, в том числе графические данные и возможные другие типы более сложных структур данных. Это актуально, например, в области астрономии и медицины, где отсрочка получения результатов моделирования на основании обработки данных должна быть строго в определенные сроки. Задержки в получении результата неприемлемы, так как теряют всякий смысл (расчеты моделирования приближающегося астероида, расчеты моделирования процессов сердечной мышцы при вводе импланта и другие).

- Эффективное индексирование. Индексирование является основополагающим для онлайн-поиска данных и поэтому важно для управления большими коллекциями документов и связанных с ними метаданных.

- Динамическая оркестровка сервисов в многосерверном и облачном контекстах. Большинство современных платформ не подходят для облачных вычислений, и обеспечение согласованности данных между различными хранилищами данных является сложной задачей.

- Параллельная обработка данных. Возможность одновременной обработки больших объемов данных очень полезна для одновременной работы с большими объемами пользователей.

**Интеллектуальный анализ данных, машинное и глубокое обучение**

Обнаружение выбросов (выявление аномалий данных) – одна из задач интеллектуального анализа данных. Это опознавание во время интеллектуального анализа данных редких данных, событий или наблюдений, которые вызывают подозрения ввиду существенного отличия от большей части данных. Компьютерные методы обнаружения выбросов можно разделить на четыре подхода: статистический подход, подход локального выброса на основе плотности, подход на основе расстояния и подход на основе отклонения. Коэффициент локальных выбросов – это алгоритм для определения локальных выбросов на основе плотности. Локальная плотность точки сравнивается с плотностью ее соседей через коэффициент локальных выбросов. Если первое значительно ниже последнего (при значении коэффициента больше единицы), точка находится в более узкой области, чем ее соседи, что говорит о том, что она является выбросом. Недостатком коэффициента локальных выбросов является то, что он работает только с числовыми данными. Еще один способ обнаружения выбросов – кластеризация. Методы кластеризации могут использоваться для идентификации кластеров одной или нескольких записей, которые удалены от других (см. таблицу 3). После группировки данных в кластеры, те данные, которые не назначены каким-либо кластерам, считаются выбросами.

Нечисловые переменные представляют некоторые проблемы, отличные от проблем числовых переменных. Некоторые инструменты, такие как деревья решений, могут обрабатывать такие значения. Другой инструмент, такой как нейронные сети, может обрабатывать только числовое представление значения. Из-за разнообразия типов баз данных некоторые базы данных могут содержать сложные объекты данных, включая временные данные, пространственные данные, данные транзакций, гипертекстовые или мультимедийные данные. Нереально ожидать, что одна система будет обрабатывать все виды данных. Следовательно, для этого существуют разные системы интеллектуального анализа данных для разных видов данных. Существует также

направление исследований, называемое сохранением конфиденциальности, которое направлено на устранение противоречий между большими данными и конфиденциальностью [Kreuter, Berg, Biemer, Decker, Lampe, Lane, O'Neil, Usher, 2015].

Из-за беспрецедентного объема данных или сложности данных часто требуется высокопроизводительный анализ данных. Высокая производительность интеллектуального анализа данных означает использование преимуществ параллельных систем управления базами данных и дополнительных процессоров для повышения производительности. Основной целью параллелизма является улучшение производительности. Есть два основных показателя улучшения производительности. Первый – это пропускная способность – количество задач, которые можно выполнить за заданный интервал времени. Второе – это время ответа – количество времени, которое требуется для выполнения одной задачи с момента ее выполнения. Две меры, как правило, количественно оцениваются по следующим показателям: увеличение и ускорение. Возникает необходимость разработки архитектуры больших данных и аналитики, которая обеспечивает: подход к управлению информацией, объединяющий все формы данных, включая структурированные, полуструктурированные и неструктурированные данные; способна обрабатывать потоки данных в пакетном режиме и в режиме реального времени; проводить высокопроизводительную аналитику в базе данных.

Таблица 3
Table 3

Алгоритмы интеллектуального анализа машинного обучения
Machine learning mining algorithms

| **Алгоритмы интеллектуального анализа машинного обучения** | **Преимущества** | **Недостатки** | **Примеры здравоохранения** |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| **Кластеризация на основе плотности** | • обрабатывает нестатические и сложные данные,<br>• обнаруживает выбросы и произвольные формы | • медленный,<br>• сложный выбор параметров,<br>• ошибки при обработке больших данных | • биомедицинская кластеризация изображений,<br>• поиск бикликов в сети |
| **Разделение на кластеры** | • простой, быстрый, полезный в обработке больших наборов данных | • высокая чувствительность к инициализации,<br>• шум и выбросы | • кластеризация депрессии, риск,<br>• прогноз реадмиссии |
| **Иерархическая кластеризация** | • возможность визуализации | • медленный,<br>• с низкой точностью, плохая визуализация для больших данных,<br>• использует огромное количество памяти | • кластеризация микрочипов,<br>• группировка по продолжительности пребывания в больнице |
| **Машина опорных векторов для классификации** | • высокая точность | • медленное обучение, вычислительно дорогой | • здоровье детей,<br>• МРТ на основе классификации |
| **Дерево решений для классификации** | • просто, легко реализовать | • ограничение, переоснащение | • МРТ, классификация мозга, медицинское прогнозирование |

Окончание табл. 3

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| **Нейронная сеть для классификации** | • обрабатывает шумные данные,<br>• обнаруживает нелинейные отношения | • медленный,<br>• с низкой точностью, модель черного ящика, вычислительно дорогой | • рак, уровень глюкозы в крови,<br>• прогнозирование, распознавание вариабельности сердечного ритма |
| **Ансамбль для классификации** | • позволяет провести прогноз, обобщение,<br>• высокая производительность | • трудно анализировать, вычислительно дорогой | • прогноз смертности,<br>• классификация, медикаментозное лечение<br>• прогноз смертности |
| **Глубокое обучение для классификации** | • глубокое обучение для классификации,<br>• обобщение, обучение,<br>• полуконтролируемое обучение,<br>• мультизадачность,<br>• большой набор данных | • трудно интерпретировать,<br>• вычислительно дорогой | • диагностика болезни Альцгеймера,<br>• регистрация<br>• МРТ головного мозга, здравоохранение,<br>• принятие решения |

Алгоритмы глубокого обучения используют огромное количество неконтролируемых данных для автоматического извлечения сложного представления. Архитектура глубокого обучения способна обобщать нелокальные и глобальные способы. Глубокое обучение позволяет извлекать представления непосредственно из неконтролируемых данных без вмешательства человека. Основным преимуществом глубокого обучения является анализ и изучение огромных объемов неконтролируемых данных, что делает его ценным инструментом для анализа больших данных, где необработанные данные в основном не имеют маркировки и не классифицируются [Zhao, 2012]. Глубокое обучение, высокопроизводительная работа с большими данными, гетерогенные вычисления повышают интеллектуальность вычислений и позволяют решить множество задач. Вопрос состоит в том, какой объем входных данных необходим для обучения и представления данных с помощью алгоритмов глубокого обучения. В таблице 3 представлены преимущества и недостатки глубокого обучения и традиционных алгоритмов интеллектуального анализа данных и машинного обучения на примере области здравоохранения. При обработке больших разнородных и разномасштабных данных традиционными методами интеллектуального анализа данных и машинного обучения существуют проблемы при обработке данных большого размера или недостаточностью данных, а также данных, не классифицированных и не контролируемых, неструктурированных, и т. д. Поэтому они имеют ограничения в аналитике больших данных.

## Заключение

Для повышения качества данных важно разработать эффективные подходы к очистке больших данных, необходимо использовать целостный, комплексный, системный подход к управлению данными, их анализу и информации. Метод главных компонент или факторный анализ часто используются для уменьшения размера данных. Неоднородность больших данных также означает одновременную работу со структурированными, полуструктурированными и неструктурированными данными. На каждом этапе анализа больших данных возникают проблемы. К ним относятся обработка в реальном времени, обработка сложных типов данных, одновременная обработка данных и т. д. Традиционные методы

интеллектуального анализа данных и машинного обучения имеют ограничения в аналитике больших данных. Глубокое обучение способно к анализу и изучению огромных объемов неконтролируемых данных; следовательно, он имеет потенциал в аналитике больших данных, где необработанные данные в основном не имеют маркировки и не классифицированы. Конфликты между аналитикой больших данных, гетерогенными вычислениями, высокопроизводительными вычислениями и глубоким обучением являются актуальной задачей обработки разнородных больших данных.

## Список литературы

1. Багутдинов Р.А. Исследование новейших информационно-коммуникативных технологий в среднем профессиональном образовании. В сборнике: Научный поиск в XXI веке. Материалы I международной научной конференции по евразийскому научному сотрудничеству. Под редакцией В.А. Должикова. 2014. С. 39–42.

2. Багутдинов Р.А. Проектирование модульной мультисенсорной системы для задач мониторинга окружающей среды на базе Arduino. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 2019. 46 (1): 173–180.

3. Багутдинов Р.А. Подход к обработке, классификации и обнаружению новых классов и аномалий в разнородных и разномасштабных потоках данных. Вестник Дагестанского государственного технического университета. Технические науки. 2018. 45 (3): 85–93.

4. Багутдинов Р.А. Разработка мультисенсорной системы для задач мониторинга и интерпретации разнородных данных. Системный администратор. 2019. 3 (196): 82–85.

5. Островский О.А. Алгоритм мероприятий по анализу ситуации при подозрении в совершении преступлений в сфере компьютерной информации с учетом специфики источников данных этой информации. Право и политика. 2018. 10: 32–37.

6. Островский О.А. Аспекты современных проблем расследования преступлений, связанных с изъятием цифровых следов и предоставлением соответствующих доказательств. Вестник Алтайской академии экономики и права. 2019. 3: 146–151.

7. Островский О.А., Шевелева И.А. Проблематика формирования и правового регулирования больших данных в исследовании информационных цифровых следов. В сборнике: Уголовное производство: процессуальная теория и криминалистическая практика. Материалы VIII Международной научно-практической конференции. Отв. редакторы М.А. Михайлов, Т.В. Омельченко. 2020. С. 57–59.

8. Kabacoff R. R. Data analysis and graphics with. Manning Publications Co.; 2015 Mar 3.

9. Zhang J, Yang X, Appelbaum D. Toward effective Big Data analysis in continuous auditing. Accounting Horizons. 2015 Jun; 29 (2): 469–76.

10. Tak PA, Gumaste SV, Kahate SA. The Challenging View of Big Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, 5 (5), May 2015, 1178–1181.

11. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Networks and Applications. 2017 Apr 1; 19 (2): 171–209.

12. Elgendy N., Elragal A. Big Data Analytics: A Literature Review Paper. P. Perner (Ed.): ICDM 2014, LNAI 8557. Springer International Publishing Switzerland, 2014, 214–227.

13. Yusuf Perwej. An Experiential Study of the Big Data. International Transaction of Electrical and Computer Engineers System, 2017, 4 (1): 14–25 (28).

14. Schotman R, Mitwalli A. Big Data for Marketing: When is Big Data the right choice? Canopy – The Open Cloud Company, 2013, p8.

15. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014: 131–40.

16. Kreuter F, Berg M, Biemer P, Decker P, Lampe C, Lane J, O'Neil C, Usher A. AAPOR Report on Big Data. Mathematica Policy Research; 2015 Feb 12.

17. Zhao Y. R. Data mining: Examples and case studies. Academic Press; 2012 Dec31.

## References

1. Bagutdinov R.A. Research of the latest information and communication technologies in secondary vocational education. In the collection: Scientific search in the XXI century. Materials of the I International

Scientific Conference on Eurasian Scientific Cooperation. Edited by V.A. Dolzhikova. 2014. S. 39–42. (in Russian)

2. Bagutdinov R.A. Designing a modular multisensor system for environmental monitoring tasks based on Arduino. Scientific Bulletin of Belgorod State University. Series: Economics. Informatics. 2019. 46 (1): 173–180. (in Russian)

3. Bagutdinov R.A. An approach to processing, classification and detection of new classes and anomalies in heterogeneous and multi-scale data streams. Bulletin of the Dagestan State Technical University. Technical science. 2018. 45 (3): 85–93. (in Russian)

4. Bagutdinov R.A. Development of a multisensor system for monitoring and interpretation of heterogeneous data. System administrator. 2019. 3 (196): 82–85. (in Russian)

5. Ostrovsky O.A. Algorithm of measures to analyze the situation in case of suspicion of committing crimes in the field of computer information, taking into account the specifics of the data sources of this information. Law and Politics. 2018. 10: 32–37. (in Russian)

6. Ostrovsky O.A. Aspects of modern problems of investigating crimes related to the removal of digital traces and the provision of relevant evidence. Bulletin of the Altai Academy of Economics and Law. 2019. 3: 146–151. (in Russian)

7. Ostrovsky O.A., Sheveleva I.A. Problems of the formation and legal regulation of big data in the study of digital information traces. In the collection: Criminal proceedings: procedural theory and forensic practice. Materials of the VIII International Scientific and Practical Conference. Resp. editors M.A. Mikhailov, T.V. Omelchenko. 2020. S. 57–59. (in Russian)

8. Kabacoff R. R. Data analysis and graphics with. Manning Publications Co.; 2015 Mar 3.

9. Zhang J, Yang X, Appelbaum D. Toward effective Big Data analysis in continuous auditing. Accounting Horizons. 2015 Jun; 29(2): 469–76.

10. Tak PA, Gumaste SV, Kahate SA. The Challenging View of Big Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, 5(5), May 2015, 1178–1181.

11. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Networks and Applications. 2017 Apr 1; 19 (2): 171–209.

12. Elgendy N., Elragal A. Big Data Analytics: A Literature Review Paper. P. Perner (Ed.): ICDM 2014, LNAI 8557. Springer International Publishing Switzerland, 2014, 214–227.

13. Yusuf Perwej. An Experiential Study of the Big Data. International Transaction of Electrical and Computer Engineers System, 2017, 4 (1): 14–25 (28).

14. Schotman R, Mitwalli A. Big Data for Marketing: When is Big Data the right choice? Canopy – The Open Cloud Company, 2013, p8.

15. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014: 131–40.

16. Kreuter F, Berg M, Biemer P, Decker P, Lampe C, Lane J, O'Neil C, Usher A. AAPOR Report on Big Data. Mathematica Policy Research; 2015 Feb 12.

17. Zhao Y. R. Data mining: Examples and case studies. Academic Press; 2012 Dec31.

## ИНФОРМАЦИЯ ОБ АВТОРАХ

**Багутдинов Равиль Анатольевич,** преподаватель электротехники и электроники, информатики и информационных технологий в профессиональной деятельности, научный руководитель, соискатель ученой степени кандидат технических наук, «Исследователь, Преподаватель-исследователь» по направлению 09.06.01 «Информатика и вычислительная техника» по специальности 05.13.01 «Системный анализ, управление и обработка информации», магистр по направлению 223200 «Техническая физика». Профессиональная образовательная организация частное учреждение «Автомобильно-дорожный колледж», Сочи, Россия

## INFORMATION ABOUT THE AUTHORS

**Ravil A. Bagutdinov,** Teacher of electrical engineering and electronics, computer science and information technology in professional activities, scientific advisor, candidate for a Ph.D. degree in science, "Researcher, Teacher-researcher" in the direction 09.06.01 "Informatics and computer technology", specialty 05.13.01 " System analysis, management and information processing ", master in the direction 223200" Technical physics". Professional educational organization Private institution "Automobile and road college". Sochi, Russia

**Саргсян Нарек Арутюнович,** студент 2 курса, по специальности 08.02.05 «Строительство и эксплуатация автомобильных дорог и аэродромов». Профессиональная образовательная организация частное учреждение «Автомобильно-дорожный колледж», Сочи, Россия

**Narek A. Sargsan,** 2nd year student, specialty 08.02.05 "Construction and operation of highways and airfields". Professional educational organization Private institution "Automobile and road college". Sochi, Russia

**Красноплахтыч Максим Александрович,** студент 2 курса по специальности 23.02.03 «Техническое обслуживание и ремонт автомобильного транспорта». Профессиональная образовательная организация частное учреждение «Автомобильно-дорожный колледж», Сочи, Россия

**Maxim A. Krasnoplakhtych,** 2nd year student, specialty 23.02.03 "Maintenance and repair of motor vehicles". Professional educational organization Private institution "Automobile and road college". Sochi, Russia