Методы интеграции, уменьшение размеров и нормализация обработки разнородных и разномасштабных данных

Р.А. Багутдинов, М.Ф. Степанов

***Аннотация*— В работе проведен анализ существующих методов обработки больших данных, которые могут быть применены к обработке разнородных и разномасштабных данных. Под разнородными данными в данной работе понимаются любые данные с высокой изменчивостью типов данных, форматов и характера происхождения. Они могут быть неоднозначными и низкого качества из-за пропущенных значений, высокой избыточности или недостоверности. Вследствие чего возникает проблема интеграции и агрегации этих данных для дальнейшей обработки или принятий конкретных решений. Особый интерес представляет получение знаний из автономных, семантически неоднородных и распределенных источников данных, ориентированных на запросы и подходов к интеграции данных. Отсутствие целостности таких данных обычно связано с недостоверными данными и неполными данными. Согласованность данных является наиболее важной проблемой для систем непрерывного аудита больших данных и связана с взаимозависимыми данными между приложениями и всей организацией. Анализ больших разнородных 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. Вследствие чего авторами рассматриваются актуальные аспекты обработки данных, выбор методов обработки данных, включая очистку данных, интеграцию данных, уменьшение размеров и нормализацию для разнородных данных и соответствующий системный и аналитический анализ, рассматривается потенциал слияния разнородных данных. В данной работе описываются некоторые преимущества и недостатки наиболее часто используемых методов обработки разнородных данных. Раскрываются проблемы обработки разнородных и разномасштабных данных. Приведены инструменты обработки больших данных, некоторые традиционные методы интеллектуального анализа данных, в том числе машинного обучения. Представлены преимущества слияния больших разнородных данных.**

***Ключевые слова*—Обработка данных, разнородные данные, разномасштабные данные, методы обработки, интеллектуальный анализ, аналитика данных, экспертные системы, системы обработки данных.**

Статья получена 19 октября 2020.

Багутдинов Равиль Анатольевич, соискатель ученой степени к.т.н. Саратовский государственный технический университет имени Ю. А. Гагарина, 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», преподаватель информатики и информационных технологий в профессиональной деятельности, научный руководитель, Профессиональная образовательная организация частное учреждение «Автомобильно-дорожный колледж», г. Сочи, Россия (e-mail: rav379@mail.ru).

Степанов Михаил Федорович, доктор технических наук, доцент, профессор кафедры «Системотехника и управление в технических системах», Институт электронной техники и приборостроения, Саратовский государственный технический университет имени Ю. А. Гагарина, Саратов, Россия (e-mail: rav379@mail.ru)

I. Введение

К разнородным данным по сути можно отнести любые данные с высокой изменчивостью типов данных, форматов и характера происхождения. Они могут быть неоднозначными и низкого качества из-за пропущенных значений, высокой избыточности или недостоверности. Вследствие чего возникает проблема интеграции и агрегации этих данных для дальнейшей обработки или принятий конкретных решений. Например, разнородные данные часто генерируются из Интернета вещей. Данные, сгенерированные из Интернета вещей, часто имеют следующие особенности:

– Разнородность. Это связанно с разнообразием устройств сбора данных.

– Разномасштабность. Данные могут быть как незначительного объема, так и большие данные. При этом должны храниться не только полученные в настоящее время данные, но и архивные данные в течение определенного периода времени.

– Эффект значительной корреляции между временем и пространством. Каждое устройство сбора данных размещается в определенном географическом местоположении, и каждый фрагмент данных имеет соответствующую отметку времени. Соотношение времени и пространства является важным свойством данных из Интернета вещей.

– Проблема эффективности и ценности данных. Из большого количества собранных данных лишь небольшая часть имеет ценность для принятия решений или дальнейшей обработки. Зачастую, чем выше объем данных, тем большое количество шумов (помех) или ненужных данных может быть изъято во время сбора и передачи данных.

Помимо прочего, можно выделить следующие типы неоднородности данных:

– Синтаксическая неоднородность, которая возникает, когда два источника данных не выражены на одном языке.

– Концептуальная неоднородность (семантическая неоднородность), также известная как

логическое несоответствие, обозначает различия в моделировании одной и той же области интересов.

– Терминологическая неоднородность, которая означает различия в именах при обращении к одним и тем же объектам из разных источников данных.

– Семиотическая неоднородность, также известная как прагматическая неоднородность, означает различную интерпретацию.

Представление данные также можно описать четырьмя уровнями.

1. «Сырые данные». Необработанные данные разных типов и из разных источников.

2. «Репрезентативность данных». Разнородные данные должны быть унифицированы. Кроме того, слишком много данных может привести к высоким когнитивным затратам на обработку данных. Этот слой преобразует отдельные атрибуты в информацию с точки зрения понятий «что, когда и где».

3. «Агрегация данных». Пространственные данные могут быть естественно представлены в виде пространственных сеток с тематическими атрибутами. Операторами обработки являются сегментация и агрегация и т. д. Агрегация облегчает визуализацию и предоставляет интуитивно понятный запрос.

4. «Обнаруженные (изъятые) и репрезентативные данные». Ситуация в местоположении таких данных характеризуется на основе пространственно-временных дескрипторов, определенных с использованием соответствующих операторов предыдущего уровня.

5. «Классификация данных». Последний шаг - это операция классификации, которая использует знание предметной области для назначения соответствующего класса в каждой ячейке данных и назначение соответствующего метода обработки данных и получения конечного результата с целью принятия дальнейших решений.

Метаданные имеют решающее значение для будущих запросов. Для реляционных таблиц и некоторых документов на расширяемом языке разметки (XML) явные определения схемы в языке структурированных запросов (SQL), определении схемы XML или определении типа документа могут быть получены непосредственно из источников и интегрированы в метамодель. Техника XML используется для перевода данных. Сложная часть - это полуструктурированные данные (такие как XML, JSON или частично структурированные файлы Excel или CSV), которые содержат неявные схемы. Поэтому компонент «Обнаружение структурных метаданных» берет на себя ответственность за обнаружение неявных метаданных (например, типов сущностей, типов отношений и ограничений) из полуструктурированных данных [1]. Вопросы управления метаданными важны. Для правильной интерпретации разнородных больших данных требуются подробные метаданные. Некоторые отчеты содержат некоторые метаданные, но для исследовательских целей требуется гораздо больше деталей, например, о конкретном датчике, используемом при сборе данных.

Разработка алгоритмов обработки больших данных сосредоточены для решения проблем, возникающих в связи с распределением этих данных, сложными и динамическими характеристиками.

Здесь можно выделить следующие этапы:

1. Неоднородные, неполные, неопределенные, разреженные, неформализованные и имеющие одновременно множество источников предварительно обрабатываются методами объединения (слияния) данных.

2. Динамические данные добываются после предварительной обработки.

3. Глобальные знания, полученные в результате локального обучения и объединения моделей, проверяются, и соответствующая информация возвращается на этап предварительной обработки. Затем модель и параметры корректируются в соответствии с обратной связью. В целом процесс обмена информацией является не только плавным развитием каждого этапа, но и одной из цели обработки больших данных.

II. Описание работы

В данной работе рассматриваются актуальные аспекты обработки данных, выбора методов обработки данных, включая очистку данных, интеграцию данных, уменьшение размеров и нормализацию для разнородных данных и соответствующий системный и аналитический анализ, рассматривается потенциал слияния разнородных данных.

Очистка данных - это процесс выявления, неполных, неточных или необоснованных данных, а затем изменения или удаления таких данных для улучшения качества данных. Например, мультимодальный характер данных приводит к проблемам высокой сложности, в том числе проблемам избыточности данных. Кроме того, существуют также проблемы с отсутствующими значениями и примесями в данных большого объема. Поскольку качество данных определяет качество информации, которое в конечном итоге будет влиять на процесс принятия решений, крайне важно разработать эффективные подходы к очистке больших данных, чтобы повысить качество данных для принятия точных и эффективных решений.

Отсутствующее значение для переменной - это значение, которое не было введено в набор данных. Пропущенные значения в переменной заменяются одним значением, например среднее значение или медиана. Однако это может привести к неточным результатам в процессе обработки, так как недооцениваются стандартные ошибки, искажается значение корреляции между переменными и могут выдаваться неверные значения в статистических тестах. Этого подхода следует избегать для большинства проблем с отсутствующими данными. Можно попытаться изучить корреляции между переменными с неизвестными и номинальными переменными. Неизвестные значения могут быть заполнены путем изучения более точных корреляций. Всякий раз, когда необходимо обработать набор данных с пропущенными значениями, можно применять следующие приемы: удаление данных с неизвестными; заполнение неизвестных значений, путем сходства между данными (сравниваются значения до и после); заполнение неизвестных значений, путем корреляции между

переменными.

База данных также может содержать нерелевантные атрибуты. Следовательно, анализ релевантности в форме корреляционного анализа и выбора подмножества атрибутов может использоваться для обнаружения атрибутов, которые не вносят вклад в задачу классификации или прогнозирования. Включение таких атрибутов может в противном случае замедлить и, возможно, ввести в заблуждение этап обучения модели. Как правило, очистка данных и интеграция данных выполняются как этап предварительной обработки. Несоответствия в именовании атрибутов или измерений могут привести к избыточности в результирующем наборе данных. Очистка данных может выполняться для обнаружения и устранения избыточностей, которые могли возникнуть в результате интеграции данных. Удаление избыточных данных часто рассматривается как основа очистки данных, а также сокращения данных.

Интеграция данных. В случае интеграции или агрегирования наборы данных сопоставляются и объединяются на основе общих переменных и атрибутов. Усовершенствованные методы обработки и анализа данных позволяют комбинировать как структурированные, так и неструктурированные данные для получения новых идей, подходов, методов; однако это требует «чистых» данных. Методы объединения данных используются для сопоставления и агрегации, для создания или улучшения представления реального положения дел, что помогает провести качественный анализ данных. Существующие методологии объединения данных среднего уровня, которые объединяют структурированные данные, в основном работают хорошо. С другой стороны, задачи объединения данных высокого уровня для объединения нескольких неструктурированных данных с различных датчиков остаются достаточно сложными.

Инструменты интеграции данных развиваются в направлении унификации структурированных и неструктурированных данных. Часто требуется структурировать неструктурированные данные и объединять разнородные источники и типы информации в единый уровень данных. Большинство платформ интеграции данных используют первичную модель интеграции, основанную на реляционных или XML-типах данных. Были предложены расширенные платформы виртуализации данных, в которых используется расширенная модель данных интеграции с возможностью хранения, чтения и записи всех типов данных в их собственном формате, таких как реляционные, многомерные, семантические данные, иерархические и индексные файлы и т. д. [2].

Интеграция разнородных источников данных является сложной задачей. Одна из причин заключается в том, что уникальные идентификаторы между записями двух разных наборов данных часто не существуют. Определение того, какие данные должны быть объединены, может быть неясным с самого начала. Работа с разнородными данными часто является итеративным процессом, в ходе которого ценность данных обнаруживается по пути, а наиболее ценные данные затем более тщательно интегрируются. Для решения подобных задач была предложена следующая интеграция: во-первых, выявление соответствий между идентичными объектами схем мультисенсорной системы, а затем уже интеграция соответствующих каталогов или кластеров данных [3, 4].

Особый интерес представляет получение знаний из автономных, семантически неоднородных и распределенных источников данных, ориентированных на запросы и подходов к интеграции данных. Для интеграции неструктурированных и структурированных данных могут использоваться следующие подходы [5]:

– Конвейеры обработки естественного языка. Может быть, непосредственно применим к проектам, которые требуют работы с неструктурированными данными.

– Распознавание и связывание сущностей. Извлечение структурированной информации из неструктурированных данных является фундаментальным шагом. Часть проблемы может быть решена с помощью методов извлечения информации, таких как распознавание сущностей, извлечение онтологий.

– Использование открытых данных для интеграции структурированных и неструктурированных данных. Объекты в открытых наборах данных можно использовать для идентификации именованных объектов (людей, организаций, мест), которые можно использовать для категоризации и организации текстового содержимого. Для связывания структурированных и неструктурированных данных можно использовать инструменты распознавания и связывания именованных объектов, например, такие как DBpedia Spotlight.

При объединении данных из разнородных мультисенсорных систем существует три источника ошибок: ошибки ввода данных, несовместимости типов данных и несовместимости в определениях объектов. Традиционно, зачастую многие предприятия используют извлечение, преобразование, загрузку (ETL) и хранилища данных (DW) для интеграции данных. Однако в последние несколько лет технология, известная как «виртуализация данных» (DV), нашла некоторое признание в качестве альтернативного решения для интеграции данных. «Виртуализация данных» - это объединенная база данных, называемая ещё составной базой данных. Виртуализация данных и стандартизация корпоративных данных обещают снизить стоимость и время внедрения интеграции данных. В отличие от DW, DV определяет очистку данных, объединение и преобразование данных программно с использованием логических представлений. DV допускает расширяемость и повторное использование, допуская цепочку логического представления. Стандартизация корпоративных данных в основном позволяет избежать несоответствия типов данных и несовместимости данных. При этом DV не является заменой DW; DV может снять определенные аналитические нагрузки с DW. Регрессионный анализ, многомерные структуры данных и анализ больших объемов данных в основном, как и ранее, не может быть без DW [6].

Озера данных (Data Lake или DL) являются новым и мощным подходом к решению задач интеграции данных, поскольку предприятия расширяют доступ к мобильным и облачным приложениям и Интернету вещей на основе различных сенсоров. Оно представляет собой огромное хранилище, в котором разные данные хранятся в «сыром», то есть неупорядоченном и необработанном виде. Видеоролики, книги, журналы, документы Word и PDF, аудиозаписи и фотографии — все это неструктурированные данные, и все они могут храниться в DL. По сути, это огромное хранилище, которое принимает любые файлы всех форматов. Источник данных тоже не имеет никакого значения. Озеро данных может принимать данные из CRM- или ERP-систем, продуктовых каталогов, банковских программ, датчиков или умных устройств — абсолютно любых систем. Озера данных являются хранилищами для большого количества разнообразных данных, как структурированных, так и неструктурированных. Такие хранилища больше подходят для менее структурированных данных, однако и при работе с ними могут возникнуть трудности, например такие как: расширение и усложнение управления метаданными над необработанными данными, извлеченными из разнородных источников данных; работа со структурными метаданными из источников данных и аннотирование данных и метаданных дополнительной информацией, чтобы избежать двусмысленности.

Поскольку структура таких данных в DL неизвестны, то без описания хранящихся данных, метаданных или моделей управления этими данными - их дальнейшая обработка будет затруднительна, т.к. все данные в таком виде будут иметь хаотическое хранение и просто представляют собой набор каких-то данных, не имеющих четкое назначение и интерес [7].

Уменьшение размеров и нормализация данных. Есть несколько причин уменьшения размерности данных. Во-первых, данные большого размера создают вычислительные проблемы. Во-вторых, высокая размерность может привести к плохим способностям обобщения алгоритма обучения в некоторых ситуациях (например, сложность выборки увеличивается экспоненциально с измерением в классификаторах ближайших соседей). Наконец, уменьшение размерности может использоваться для нахождения значимой структуры данных, интерпретируемости данных и целей иллюстрации [8].

Выбор подмножества функций является широко известной задачей интеллектуального анализа данных и машинного обучения. Генетические алгоритмы - это часто используемые алгоритмы для задач выбора подмножества объектов. Уменьшение размерности, обеспечиваемое процессом подмножества функций, может обеспечить несколько преимуществ: более быстрое введение окончательной модели классификации; улучшение понятности окончательной модели классификации; повышение точности классификации.

Методы выбора признаков можно разделить на два подхода: ранжирование признаков и выбор поднабора. При первом подходе объекты ранжируются по некоторым критериям, а затем выбираются объекты выше определенного порога. Во втором подходе каждый ищет пространство подмножеств признаков для оптимального подмножества. Более того, второй подход можно разделить на три части: подходы фильтра - сначала выбирают признаки, а затем используют это подмножество для выполнения алгоритма классификации; встроенные подходы - выбор признаков происходит как часть алгоритма классификации; и алгоритм определения по набору данных используется для определения лучших характеристик.

В наборе данных с большим количеством переменных, как правило, существует много совпадений в информации. Один простой способ найти избыточность - это проверить корреляционную матрицу, полученную с помощью корреляционного анализа. Далее факторный анализ - это метод уменьшения размерности, для понимания основных причин корреляции между группой переменных. Факторный анализ может быть использован для уменьшения количества переменных и выявления структуры в отношениях между переменными. Поэтому факторный анализ часто используется как метод определения структуры или сокращения данных [9]. Кроме того, метод главных компонент также полезен, когда имеются данные о большом количестве переменных и, возможно, имеется некоторая избыточность в этих переменных. В этой ситуации избыточность означает, что некоторые переменные коррелируют друг с другом. Метод главных компонент очень быстрый, эффективный, простой и широко используемый. Выделим некоторые моменты этого метода:

Предварительная обработка. Изучение сложных моделей многомерных данных часто происходит очень медленно и также подвержено переобучению. Количество параметров в модели обычно экспоненциально по количеству измерений. С помощью метода главных компонент можно также выделить элементы, которые перебалансирует вес данных, чтобы в некоторых случаях повысить производительность вычислений.

Моделирование. Метод главных компонент иногда используется как целая модель, например, предварительное распределение для новых данных.

Сжатие. Метод главных компонент (МГК) может использоваться для сжатия данных, заменяя данные на их низкоразмерное представление. Основные этапы использования МГК или факторного анализа заключаются в следующем: подготовка данных, таких, как проверка данных на наличие пропущенных значений; выбор фактор-модели, решив, лучше ли подходит МГК или факторный анализ для целей исследования, и выбрав конкретный метод факторинга (например, максимальную вероятность), если выбран подход факторного анализа; решить, сколько компонентов / факторов извлечь; и извлечь компоненты / факторы. Что касается выбора количества компонентов для извлечения, существует несколько критериев для определения количества компонентов. Они включают в себя: основание количества компонентов на предыдущем опыте и теории; выбор количества

компонентов, необходимых для учета некоторой пороговой совокупной величины дисперсии в переменных (например, 80%); выбор количества сохраняемых компонентов путем изучения собственных значений матрицы корреляции среди переменных [10].

III. Заключение

Некоторые алгоритмы требуют, чтобы данные были нормализованы (стандартизированы), прежде чем алгоритм может быть эффективно реализован. Нормализация (или стандартизация) означает замену каждой исходной переменной стандартизированной версией переменной, которая имеет единичную дисперсию. Эффект этой нормализации (стандартизации) состоит в том, чтобы дать всем переменным равную важность с точки зрения изменчивости, при этом данные часто нормализуются перед выполнением МГК.

Библиография

[1] *Багутдинов Р.А.* Разработка мультисенсорной системы для задач мониторинга и интерпретации разнородных данных // Системный администратор. 2019. №3 (196). С. 82-85.

[2] *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. Т. 45. №3. С. 85-93. https://doi.org/10.21822/2073-6185-2018-45-3-85-93

[3] *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. С. 32-37. https://doi.org/10.7256/2454-0706.2018.10.22879

[4] *Островский О.А.* Специфика алгоритма назначения ситуационных экспертиз // Судебно-медицинская экспертиза. 2019. Т. 62. №2. С. 48-51. https://doi.org/10.17116/sudmed20196202148

[5] *Curry E, Kikiras P, Freitas A.* Big Data Technical Working Groups // White Paper, BIG Consortium, 2012.

[6] *Chen M, Mao S, Liu Y.* Big data: A survey // Mobile Networks and Applications. 2014 Apr 1; 19(2): 171-209.

[7] *Daniel D.* Gutierrez, InsideBIGDATA Guide to Big Data for Finance // White Paper, DELL and intel, Whitepaper, 2015, 1-14.

[8] *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.

[9] *Harrington P.* Machine learning in action // Greenwich, CT: Manning; 2012 Apr 16.

[10] *Hai R, Geisler S, Quix C.* Constance: An intelligent data lake system. // In Proceedings of the 2016 International Conference on Management of Data 2016 Jun 26 (pp. 2097-2100). ACM.

[11] *Jaseena K.U, David J.M.* Issues, challenges, and solutions: big data mining. // NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014: 131-40.

[12] *Kabacoff R. R.* Data analysis and graphics with // Manning Publications Co.; 2015 Mar 3.

[13] *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.

[14] *Najafabadi M.M., Villanustre F., Khoshgoftaar T.M., Seliya N., Wald R., Muharemagic E.* Deep learning applications and challenges in big data analytics // Journal of Big Data. 2015 Feb 24; 2(1): 1.

[15] *Pullokkaran L.J.* Analysis of data virtualization & enterprise data standardization in business intelligence // Doctoral dissertation, Massachusetts Institute of Technology, 2013.

[16] *Rudin C., Dunson D., Irizarry R., Ji H., Laber E., Leek J., & Wasserman L.* Discovery with data: Leveraging statistics with computer science to transform science and society. July 2, 2014, 1-27.

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

[18] *Stein B, Morrison A.* The enterprise data lake: Better integration and deeper analytics // PwC Technology Forecast: Rethinking integration. 2014(1), 1-9.

[19] *Shalev-Shwartz S, Ben-David S.* Understanding machine learning: From theory to algorithms // Cambridge university press; 2014 May 19.

[20] *Tak P.A, Gumaste S.V, Kahate S.A.* The Challenging View of Big Data Mining // International Journal of Advanced Research in Computer Science and Software Engineering, 5(5), May 2015, 1178-1181.

[21] *Vina A.* Data Virtualization Goes Mainstream // White Paper, Denodo Technologies, Inc, USA, 2015, 1-18.

[22] *Yusuf Perwej.* An Experiential Study of the Big Data // International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 1, 14-25 (28)

[23] *Zhang J, Yang X, Appelbaum D.* Toward effective Big Data analysis in continuous auditing // Accounting Horizons. 2015 Jun; 29(2):469-76.

[24] *Zhao Y. R.* Data mining: Examples and case studies // Academic Press; 2012 Dec31.

Багутдинов Равиль Анатольевич родился в г. Караганде 26.04.1985, закончил с отличием Карагандинский высший политехнический колледж в 2006 году по специальности «Сети связи и системы коммутаций», получил степень бакалавра по специальности «Радиотехника, электроника и телекоммуникации» в Карагандинском государственном техническом университете в 2009 году, получил степень магистра по специальности 223200 «Техническая физика» в Национальном исследовательском Томском государственном университете в 2013 году, закончил с отличием аспирантуру в Национальном исследовательском Томском политехническом университете, получил квалификацию «Исследователь. Преподаватель-исследователь» по направлению 09.06.01 «Информатика и вычислительная техника» по специальности 05.13.01 «Системный анализ, управление и обработка информации», кандидатские экзамены сданы на отлично, готовит к защите диссертацию на соискание ученой степени кандидата технических наук. Соискатель ученой степени к.т.н. Саратовский государственный технический университет имени Ю. А. Гагарина.

Область научных интересов: теоретические основы и методы системного анализа, оптимизации, управления, принятия решений и обработки информации; методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических системах; визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации; вопросы технического зрения и телекоммуникации в робототехнике для космической и военной отрасли; вопросы модернизации и оптимизации системы образования.

Степанов Михаил Федорович, доктор технических наук, доцент, профессор, зав. кафедрой «Системотехника и управление в технических системах», Институт электронной техники и приборостроения, Саратовский государственный технический университет имени Ю. А. Гагарина, Саратов, Россия (e-mail: rav379@mail.ru)

Methods of integration, reduction of sizes and normalization of processing of heterogeneous and multi-scale data

R. A. Bagutdinov, M. F. Stepanov

Abstract— The paper analyzes the existing methods for processing big data, which can be applied to the processing of heterogeneous and multi-scale data. In this work, heterogeneous data is understood as any data with high variability of data types, formats and nature of origin. They can be ambiguous and of poor quality due to missing values, high redundancy, or unreliability. As a result, there is a problem of integrating and aggregating this data for further processing or making specific decisions. Of particular interest is the acquisition of knowledge from autonomous, semantically heterogeneous and distributed data sources, query-oriented and approaches to data integration. The lack of integrity of such data is usually associated with invalid data and incomplete data. Data consistency is the most critical issue in continuous auditing systems for big data and relates to interdependent data between applications and the entire organization. Analyzing large, heterogeneous data can be problematic because it often involves collecting and storing mixed data based on different patterns or rules. The context of the data and their description play an important role here. As a result, the authors consider relevant aspects of data processing, the choice of data processing methods, including data cleansing, data integration, size reduction and normalization for heterogeneous data and the corresponding system and analytical analysis, the potential for fusion of heterogeneous data is considered. This paper describes some of the advantages and disadvantages of the most commonly used methods for processing heterogeneous data. The problems of processing heterogeneous and different-scale data are revealed. The tools for processing big data, some traditional methods of data mining, including machine learning are presented.

***Keywords* — Data processing, heterogeneous data, multi-scale data, processing methods, data mining, data analytics, expert systems, data processing systems.**

REFERENCES

[1] Bagutdinov R.A. Development of a multisensor system for monitoring and interpretation of heterogeneous data // System Administrator. 2019.No. 3 (196). S. 82-85.

[2] 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.Vol. 45. No. 3. S. 85-93. https://doi.org/10.21822/2073-6185-2018-45-3-85-93

[3] 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. No. 10. S. 32-37. https://doi.org/10.7256/2454-0706.2018.10.22879

[4] Ostrovsky O.A. Specificity of the algorithm for assigning situational examinations // Forensic medical examination. 2019.Vol. 62.No.2. S. 48-51. https://doi.org/10.17116/sudmed20196202148

[5] Curry E, Kikiras P, Freitas A. Big Data Technical Working Groups // White Paper, BIG Consortium, 2012.

[6] Chen M, Mao S, Liu Y. Big data: A survey // Mobile Networks and Applications. 2014 Apr 1; 19(2): 171-209.

[7] Daniel D. Gutierrez, InsideBIGDATA Guide to Big Data for Finance // White Paper, DELL and intel, Whitepaper, 2015, 1-14.

[8] 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.

[9] Harrington P. Machine learning in action // Greenwich, CT: Manning; 2012 Apr 16.

[10] Hai R, Geisler S, Quix C. Constance: An intelligent data lake system. // In Proceedings of the 2016 International Conference on Management of Data 2016 Jun 26 (pp. 2097-2100). ACM.

[11] Jaseena K.U, David J.M. Issues, challenges, and solutions: big data mining. // NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014: 131-40.

[12] Kabacoff R. R. Data analysis and graphics with // Manning Publications Co.; 2015 Mar 3.

[13] 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.

[14] Najafabadi M.M., Villanustre F., Khoshgoftaar T.M., Seliya N., Wald R., Muharemagic E. Deep learning applications and challenges in big data analytics // Journal of Big Data. 2015 Feb 24; 2(1): 1.

[15] Pullokkaran L.J. Analysis of data virtualization & enterprise data standardization in business intelligence // Doctoral dissertation, Massachusetts Institute of Technology, 2013.

[16] Rudin C., Dunson D., Irizarry R., Ji H., Laber E., Leek J., & Wasserman L. Discovery with data: Leveraging statistics with computer science to transform science and society. July 2, 2014, 1-27.

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

[18] Stein B, Morrison A. The enterprise data lake: Better integration and deeper analytics // PwC Technology Forecast: Rethinking integration. 2014(1), 1-9.

[19] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms // Cambridge university press; 2014 May 19.

[20] Tak P.A, Gumaste S.V, Kahate S.A. The Challenging View of Big Data Mining // International Journal of Advanced Research in Computer Science and Software Engineering, 5(5), May 2015, 1178-1181.

[21] Vina A. Data Virtualization Goes Mainstream // White Paper, Denodo Technologies, Inc, USA, 2015, 1-18.

[22] Yusuf Perwej. An Experiential Study of the Big Data // International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 1, 14-25 (28)

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

[24] Zhao Y. R. Data mining: Examples and case studies // Academic Press; 2012 Dec31.