МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ / MATHEMATICAL SOFTWARE FOR COMPUTERS, COMPLEXES AND COMPUTER NETWORKS

DOI: https://doi.org/10.23670/IRJ.2022.123.28

РАЗРАБОТКА МОДУЛЯ, РЕАЛИЗУЮЩЕГО МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ РАЗЛИЧНЫХ ТИПОВ ПОТОКА ДАННЫХ

Научная статья

Тапе Ж.[1, *], Погуда А.А.[2]
[1] ORCID : 0000-0002-7438-5279;
[1, 2] Национальный исследовательский Томский государственный университет, Томск, Российская Федерация

* Корреспондирующий автор (jeanmax.habib[at]mail.ru)

Аннотация

В настоящее время объем данных, генерируемых машинами и человеческими взаимодействиями, быстро растет, и технологии развиваются, пытаясь решить эту проблему. Хотя большие данные широко обсуждаются на теоретическом уровне, существует ряд трудностей при их обработке.

Целью данной работы является разработка модуля, который позволит классифицировать поток данных и затем обрабатывать его, принимая во внимание определенные параметры, такие как: типы файлов в соответствии с расширением типа, дата, имя и размер файла, используя в доказательство определенные методы и алгоритмы.

Очевидно, что этот модуль позволит легче и быстрее обрабатывать и устранять определенные трудности, связанные со структурой больших данных.

Ключевые слова: большие данные, структуры больших данных, алгоритм, набор данных.

DEVELOPMENT OF MODULE IMPLEMENTING METHODS AND ALGORITHMS FOR PROCESSING VARIOUS TYPES OF DATA FLOW

Research article

Tape J.[1, *], Poguda A.A.[2]
[1] ORCID : 0000-0002-7438-5279;
[1, 2] National Research Tomsk State University, Tomsk, Russian Federation

* Corresponding author (jeanmax.habib[at]mail.ru)

Abstract

The amount of data generated by machines and human interactions is now growing rapidly, and technologies are evolving trying to solve this problem. Although big data is widely discussed on a theoretical level, there are a number of difficulties in its processing.

The aim of this work is to develop a module that will classify a data flow and then process it, taking into account certain parameters, such as: file types according to type extension, date, file name and size, using certain methods and algorithms as proof.

Obviously, this module will allow easier and faster processing as well eliminate certain difficulties associated with the structure of big data.

Keywords: Big data, big data structures, algorithm, data set.

Введение

На сегодняшний день, одним из активно развивающихся направлений в области информационных технологий является технология Больших данных (Big Data). В последние годы Большие данные являются общепризнанным признаком экономического и технологического развития. Исследования консалтинговой компании «Gartner» прогнозируют, что технология Больших данных окажет существенное влияние на информационные технологии в производстве, здравоохранении, торговле, государственном управлении и в других отраслях, которые используют большой поток информации.

Big Data – это структурированные или неструктурированные массивы данных большого объема. Их обрабатывают при помощи специальных автоматизированных инструментов, чтобы использовать для статистики, анализа, прогнозов и принятия решений [4].

Термин «Большие данные» обычно относится к наборам данных, которые превышают возможности обычно используемых программных средств по сбору, хранению, управлению и обработке данных в допустимые сроки. В целом, Большие данные можно объяснить в терминах трёх «V»: Volume (объем данных), Velocity (скорость данных) и Variety (диапазон типов и источников данных), которые требуют новой высокопроизводительной обработки [3]. Реализация систем Больших Данных опирается на такие революционные технологии, как облачные вычисления, Интернет вещей и аналитика данных. Поскольку все больше систем используют Большие Данные в различных секторах, таких как здравоохранение, правительство, сельское хозяйство, оборона и образование, в областях их применения были достигнуты прорывы за счёт инноваций и роста. Эти системы представляют собой крупные, долгосрочные инвестиции, требующие значительных финансовых обязательств и масштабного развёртывания

программного обеспечения и систем. Для их обработки было разработано несколько технологий, которые классифицируются по концепциям обработки данных. Из собранной информации необходимо извлечь и проанализировать большое количество контента, чтобы удовлетворить потребности в знаниях различных бизнес-организаций, политических партий и научно-исследовательских отделов. Процесс начинается с извлечения информации, которая может поступать из различных источников, таких как базы данных, веб-сайты, документы или системы управления контентом. Вслед за хроникой их нужно фильтровать и оптимизировать. Только релевантная информация должна регистрироваться методами, исключающими ненужные данные. Для поддержки этой работы используются специальные инструменты, например, ETL. Метод ETL обычно объединяет данные из нескольких систем, а затем загружает их в хранилище данных [1]. Современный объем данных, которыми управляют наши системы, превысил возможности обработки традиционных систем [2], и это также относится к добыче данных. Появление новых технологий и услуг (таких как облачные вычисления), а также снижение цен на компьютерное оборудование приводит к постоянному увеличению объема информации в интернете. Это явление, безусловно, представляет собой «большой» вызов для сообщества специалистов по анализу данных.

Однако существуют некоторые проблемы, такие как масштабируемость, интеграция, отказоустойчивость, своевременность, согласованность, неоднородность и неполнота, балансировка нагрузки, проблемы конфиденциальности и точность [5], [6], которые возникают из-за природы потоков больших данных, с которыми необходимо иметь дело.

Методы и принципы исследования

Гетерогенная структура, различная размерность и разнообразие представления данных также относятся к этому вопросу. Просто вспомните старые приложения, выполняющие запись данных: различные программные реализации приведут к различным схемам и протоколам [18]. Для больших компаний статистика и анализ данных лежат в основе ведения бизнеса на крупных рынках, но аналитика стала намного более востребованной с развитием телекоммуникаций и намного более эффективной благодаря наличию мощных вычислительных машин. Один из экспертов Big Data Фрэнк Акито, считает, что сильнейшим фактором расширения спектра применения Big Data является Интернет [7].

Существует множество методов классификации, которые используют различный математический аппарат и различные подходы при реализации [9], [10], [11], [12]. Однако эффективность этих методов зависит от конкретной решаемой задачи. Несмотря на то, что последнее десятилетие коммерческие компании занимаются проблемой повышения качества машинного обучения, на сегодняшний день не существует методов, которые могли бы однозначно эффективно решить задачу классификации. Можно выделить следующие типы методов классификации: вероятностные, метрические, логические, линейные, логическая регрессия. Обобщенно опишем некоторые из них, указывая преимущества и недостатки каждого из них.

Методы и алгоритмы анализа больших данных

Существует множество разнообразных методик анализа массивов данных, в основе которых лежит инструментарий, [8] заимствованный из статистики и информатики. В данном списке отражены наиболее востребованные в различных отраслях подходы. При этом следует понимать, что исследователи продолжают работать над созданием новых методик и совершенствованием существующих. Кроме того, некоторые из перечисленных них методик вовсе не обязательно применимы исключительно к большим данным и могут с успехом использоваться для меньших по объему массивов (например, A/B-тестирование, регрессионный анализ). Безусловно, чем более объемный и диверсифицируемый массив подвергается анализу, тем более точные и релевантные данные удаётся получить на выходе.

Машинное обучение − узкоспециализированная область знаний, входящая в состав основных источников технологий и методов, применяемых в областях больших данных и Интернета вещей, которая изучает и разрабатывает алгоритмы автоматизированного извлечения знаний из сырого набора данных, обучения программных систем на основе полученных данных, генерации прогнозных и/или предписывающих рекомендаций, распознавания образов и т.п.

Алгоритмы машинного обучения: Линейная и логистическая регрессия; SVM; Решающие деревья; Random forest; AdaBoost; Градиентный бустинг; Нейросети; K-means; EM-алгоритм; Авторегрессии; Self-organizing maps.

Основные методы обработки больших данных включают в себя, смешение и интеграцию разнородных данных, таких как

- цифровая обработка сигналов и обработка естественного языка;
- машинное обучение, включая искусственные нейронные сети, сетевой анализ, методы оптимизации и генетические алгоритмы;
- распознавание образов;
- прогнозную аналитику;
- имитационное моделирование;
- пространственный и статистический анализ;
- визуализацию аналитических данных — рисунки, графики, диаграммы, таблицы.

Анализ больших данных может быть охарактеризован по следующим параметрам:

1 - Объем, т.е. количество генерируемых данных. От этого показателя зависит, может ли определённый массив данных считаться большими данными или нет. Данных хранятся SQL-серверах в облачной среде.

2 - Многообразие, т.е. категория, к которой принадлежат большие данные. Знание такой принадлежности позволяет аналитикам наиболее эффективно работать с информацией.

3 - Скорость, т.е. скорость генерирования или обработки данных с целью осуществления поставленных целей.

4 - Изменчивость, т.е. нестабильность данных во времени.

5 - Достоверность, т.е. качество собранных данных, от которого зависит точность анализа.

6 - Сложность, т.е. трудоёмкость процесса корреляции и построения взаимосвязей между данными.

Как мы видели в предыдущих разделах, было предложено множество алгоритмов и методик для решения проблем интеграции Больших Данных, но многие элементы не учитываются в этих предложениях. В отношении некоторых проблем, связанных с Большими Данными, из которых эта статья посвящена решению проблем, связанных с обработкой различных данных.

Постановка задачи

ИТ-индустрия давно создала методологию и инструменты для работы со структурированными данными - это реляционная модель данных и системы управления базами данных. Но современной тенденцией является необходимость работы с широким спектром данных, и это та область, где предыдущие подходы работают плохо или не работают вообще из-за несовместимости методов и алгоритмов.

Именно эта потребность требует нового способа работы с данными, и модель Больших Данных становится все более популярной. Задача разработчиков и учёных в области Больших Данных - найти программное и техническое решение, которое можно легко интегрировать в существующую инфраструктуру центров обработки данных и обеспечить этапы обработки информации.

Основные результаты

Случай алгоритма распознавания образов объектов, встроенных в модули

Для выполнения операции сортировки деталей на конвейере предлагается использовать алгоритм распознавания на основе анализа контуров — границ, изображений объектов. В этих алгоритмах в этапе представления и описания границы объектов используется метод Фурье. Для вычисления фурье-дескрипторов контур границы объекта представляется в виде массива комплексных чисел f(k) =*x*(k)+*iy*(k), k=0,1, …, N = −1. Выражение дискретного преобразования Фурье для массива f(k) задается выражением [19].

$$F(u) = \frac{1}{N}\sum_{k=0}^{N-1} f(k)\exp\left(-i\frac{2\pi uk}{N}\right), u = 0, 1, \ldots, N-1 \quad (1)$$

Комплексные коэффициенты F(u) = F_u, определяемые выражением (1), называются фурье-дескрипторами границы. При формировании вектора признаков используют модули фурье-дескрипторов с положительными и отрицательными индексами u = ±1, ±2,…, ±L/2, причем $L \le N-1$. Для обеспечения инвариантности признаков к сдвигу, повороту и изменению масштаба выполняют нормировку дескрипторов на модуль дескриптора с индексом u =1. Вектор признаков X имеет вид

$$\mathrm{X} = \left(\frac{|F_{-L/2}|}{|F_1|}, \ldots, \frac{|F_{-2}|}{|F_1|}, \ldots, \frac{|F_2|}{|F_1|}, \ldots, \frac{|F_{L/2}|}{|F_1|}\right)^T \quad (2)$$

Значение параметра L определяет размерность признакового пространства $K = L-2$.

На этапе распознавания выбран простейший классификатор, основанный на минимизации евклидова расстояния между вектором $\mathbf{X}^r$ признаков распознаваемого объекта ω_r и векторами $\mathbf{X}^m$, $m = \overline{1, M}$. признаков эталонных объектов, образующих алфавит классов [19]. В данном случае число классов равно $\boldsymbol{M}$. Евклидово расстояние определяется по формуле $\rho_{rm} = \|\mathrm{X}^r - \mathrm{X}^m\| = \sqrt{\sum_{i=1}^{K}\left(X_i^r - X_i^m\right)^2}$.

Решение о принадлежности объекта ω_r к некоторому классу Ω_{m^*} .

Реализация

Решение перечисленных задач может лежать в следующем:

Эта работа состоит из двух основных частей:

- этап подготовки данных: это этап группировки данных в соответствии с их типом или расширением в каталоге с помощью библиотеки shutil. Классификация рассматривается как метод контролируемого обучения в машинном обучении, относящийся также к проблеме прогностического моделирования, когда для данного примера предсказывается метка класса [17]. Математически, он отображает функцию (f) от входных переменных (X) на выходные переменные (Y) в виде цели, метки или категории. Чтобы предсказать класс заданных точек данных, это может быть выполнено на структурированных или неструктурированных данных. Например, обнаружение спама, такого как «спам» и «не спам» в почтовых службах может быть проблемой классификации.

— этап фактического лечения - это этап избирательного выбора лечения в соответствии с нашими потребностями, по принципу и с учетом адекватных методов и алгоритмов.

Шаг 0: создайте набор данных, содержащий несколько типов файлов.

Шаг 1: загрузка набора данных и предварительная обработка данных

Шаг 2: классификация данных по группам в соответствии с расширением

Шаг 3: преобразование и выбор метода обработки данных.

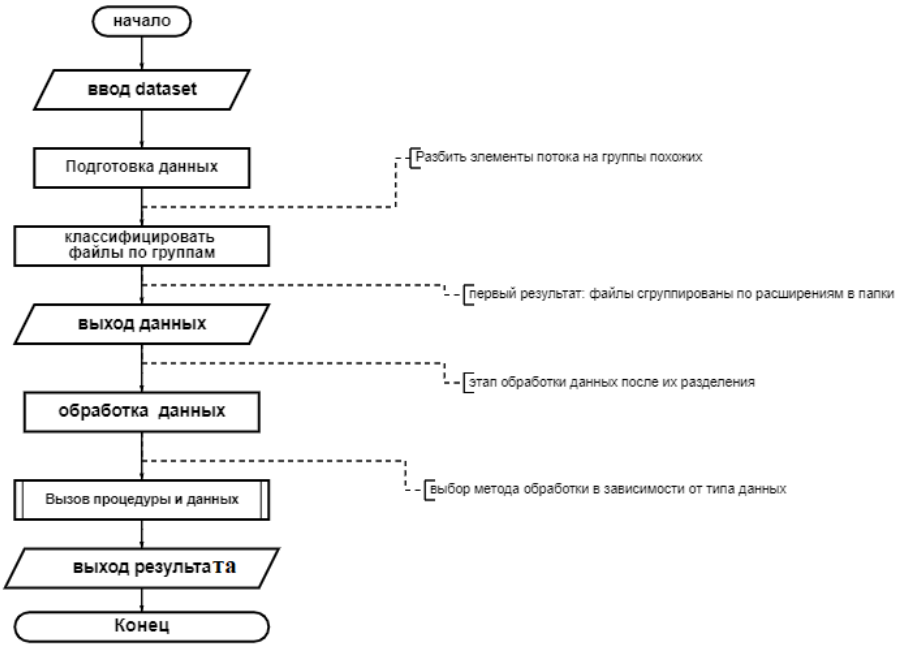

Рисунок 1 - Общая схема предлагаемого алгоритма

DOI: https://doi.org/10.23670/IRJ.2022.123.28.1

Библиотека python:

Tkinter ― для создания графического интерфейса;

shutil.move ― Эта функция используется для перемещения файла и каталога из одного каталога в другой и удаления его из предыдущего каталога[13]. Эта функция позволит нам классифицировать файлы по расширению в отдельной папке

```
def filTransprocess():
   list_of_files = os.listdir(path)
   for file in list_of_files:
      name, ext = os.path.splitext(file)
      ext = ext[1:]
      if ext == '':
         continue
      if os.path.exists(path+'/'+ext):
         shutil.move(path+'/'+file,path+'/'+ext+'/'+file)
      else:
         os.makedirs(path+'/'+ext)
         shutil.move(path+'/'+file,path+'/'+ext+'/'+file)
         return name
```

OS поставляется со стандартными служебными модулями Python [14].

KImage даёт дополнительную функциональность для обработки группы изображений как одного изображения [15].

Matplotlib - это основная библиотека для построения научных графиков в Python. [16]

NumPy В результате в процессе выполнения основных NumPy-операций (срезов, масок, индексирования) есть возможность менять пиксельные значения изображения [16].

Практический этап и результат работы

Рисунок 2 - Интерфейс программы

DOI: https://doi.org/10.23670/IRJ.2022.123.28.2

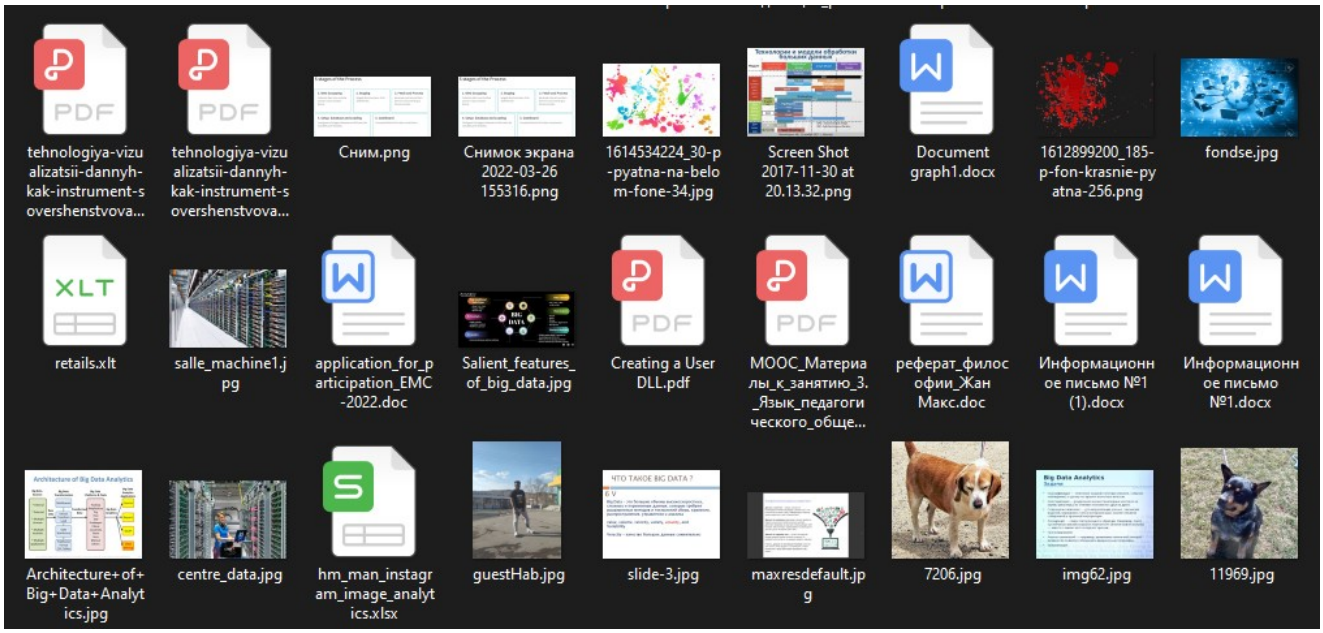

Рисунок 3 - Создать папку и импортировать несколько файлов с разными типами расширений, которая будет рассматриваться как набор данных(dataset)

DOI: https://doi.org/10.23670/IRJ.2022.123.28.3

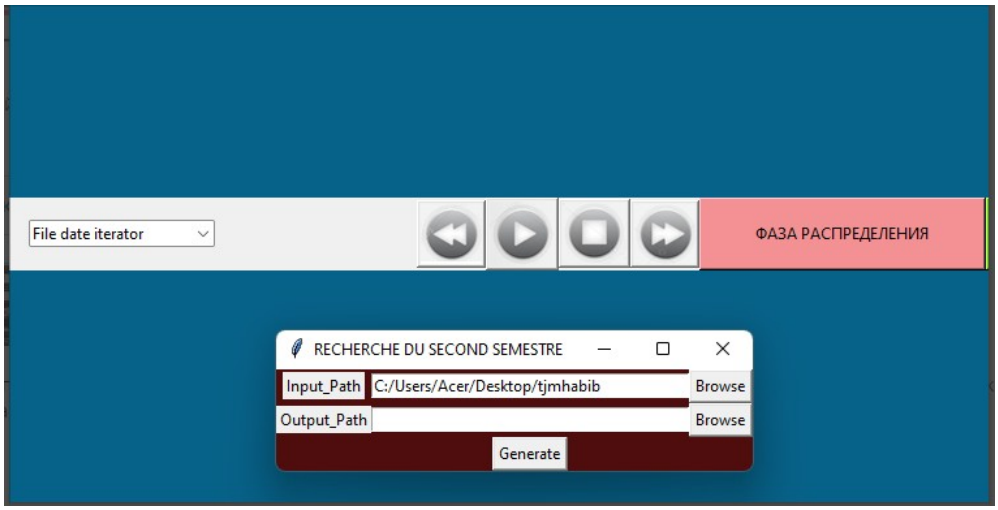

Рисунок 4 - Импорт набора данных в программное обеспечение и группировка файлов по типам расширений

DOI: https://doi.org/10.23670/IRJ.2022.123.28.4

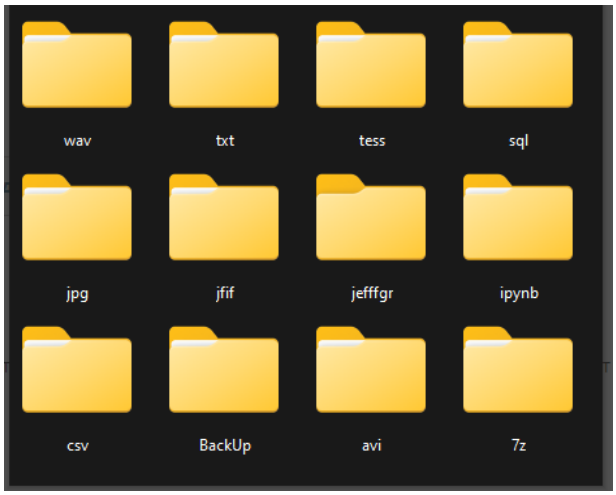

Рисунок 5 - Результат группировки файлов по типу расширения

DOI: https://doi.org/10.23670/IRJ.2022.123.28.5

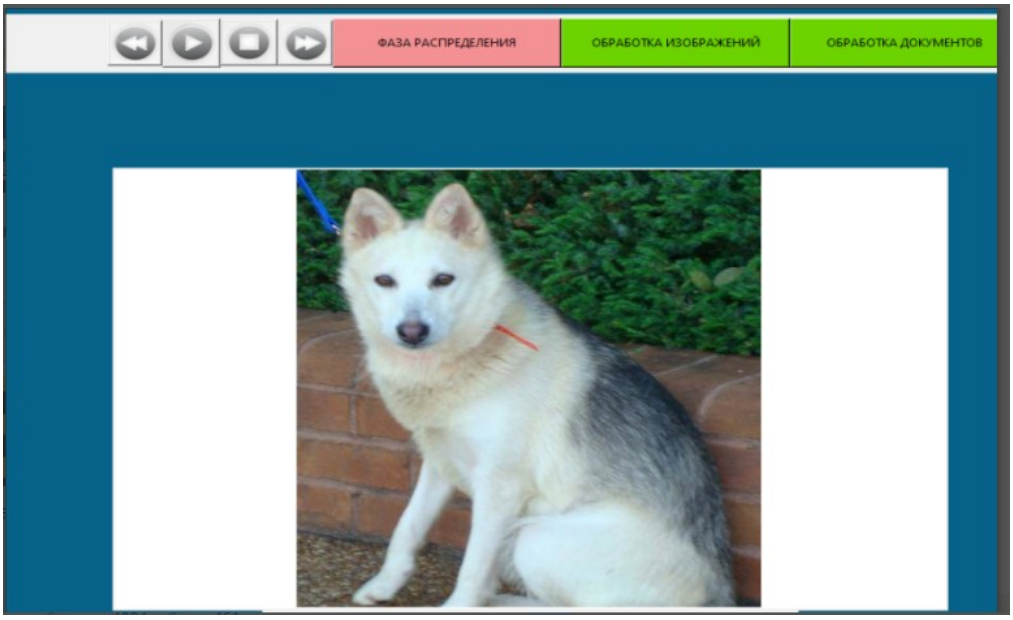

Рисунок 6 - Выбор обработки изображений

DOI: https://doi.org/10.23670/IRJ.2022.123.28.6

Рисунок 7 - Обработка изображений, фильтрация и негативное изображение

DOI: https://doi.org/10.23670/IRJ.2022.123.28.7

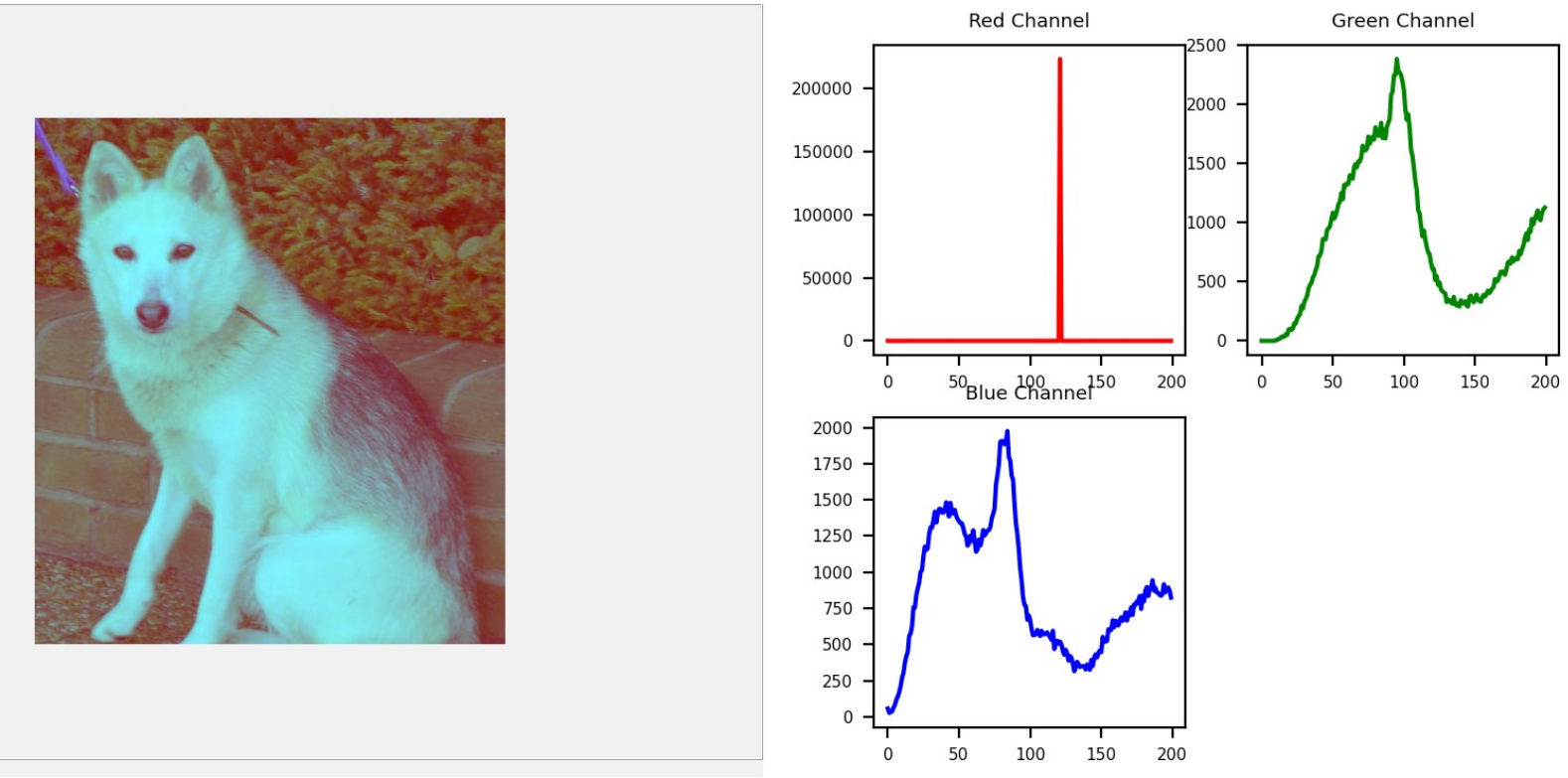

Рисунок 8 - Обработка изображений, фильтрация и гистограмма

DOI: https://doi.org/10.23670/IRJ.2022.123.28.8

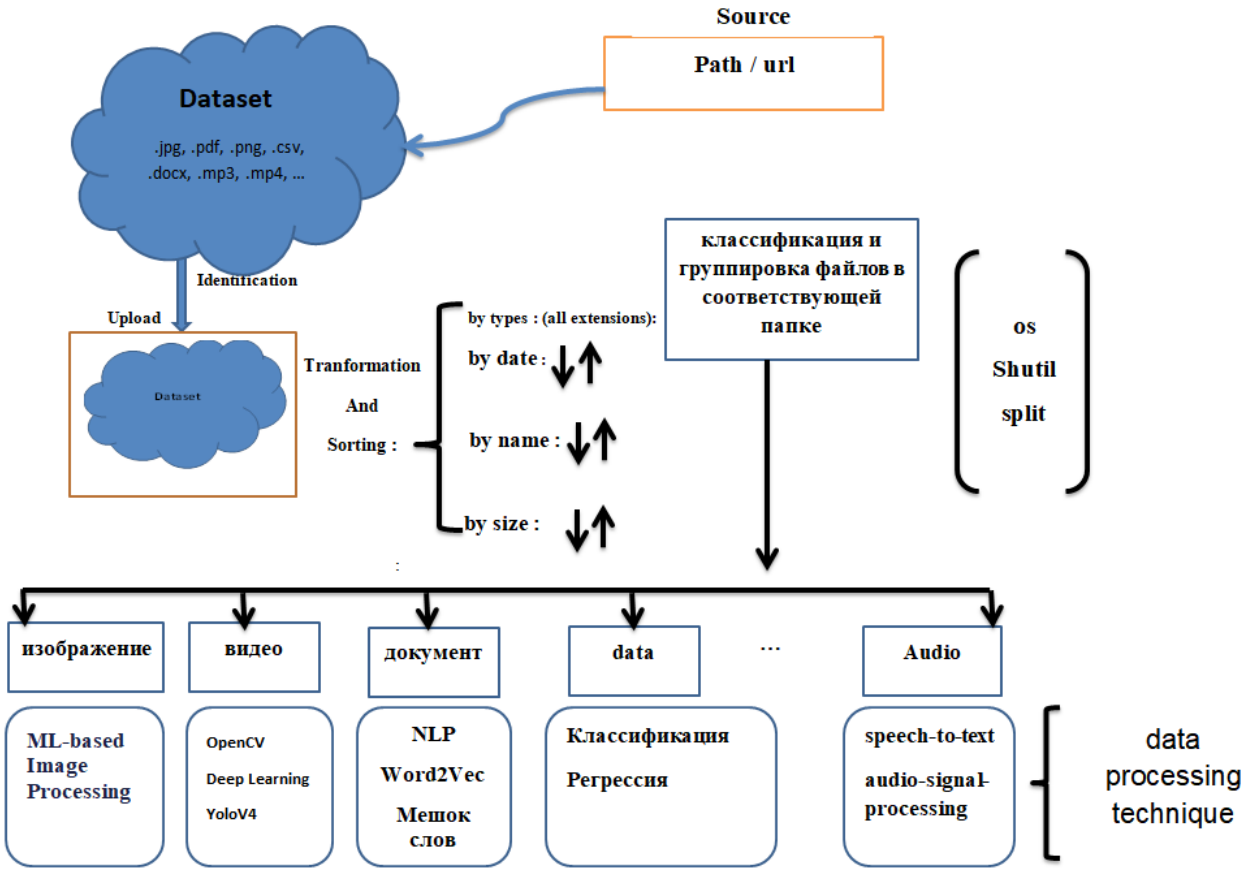

Рисунок 9 - Схема описания работы модуля анализа и обработки файлов

DOI: https://doi.org/10.23670/IRJ.2022.123.28.9

Специфика методов и алгоритмов, разработанных в модуле, позволяет интегрировать и проводить совместный анализ для выполнения конкретных функций обработки:

— OCR: Оптическое распознавание символов работает путем разделения изображения текстового символа на участки и различения пустых и непустых областей. В зависимости от шрифта или сценария, используемого для письма, контрольная сумма полученной матрицы впоследствии маркируется (первоначально человеком) как соответствующая символу на изображении

— распознавание объектов на фотографиях и видео.

— преобразование документов

— преобразование аудио в текст и наоборот

—обработка изображений (фильтрация, обнаружение разрывов, математическая морфология...)

—анализ изображений (связанные компоненты, корректировка примитивов...)

—графический интерфейс (отображение изображений, видео, управление событиями...)

— расчет гистограммы оттенков серого или цветовых гистограмм.

— сглаживание, фильтрация.

— воспроизведение, запись и просмотр видео (из файла или камеры)

— обнаружение прямых, сегментов и окружностей с помощью преобразования Хью

— распознавание лиц методом Виолы и Джонса

— обнаружение движения, история движения

— обнаружение точек интереса

— зрение (калибровка камеры, стереовидение, поиск ассоциаций...)

Учитывая несовместимость методов и алгоритмов при обработке больших данных из-за структур данных, этот комплекс позволяет использовать несколько типов данных.

Обсуждение

Для решения проблем интеграции больших данных было предложено множество алгоритмов и методик, но многие элементы не учитываются в этих предложениях. Прежде всего, все предложения предполагают, что данные хорошо сформированы и проведена их предварительная обработка (извлечение, преобразование). Кроме того, единственным типом рассматриваемых неструктурированных данных является текст, но часто встречаются и другие типы, например,

посты в социальных сетях могут содержать видео, аудио, изображения, карты и другие типы. Кроме того, при интеграции данных из социальных сетей (блоги, твиты, посты ...), данные обычно очень низкого качества, поскольку они предоставляются обычными пользователями, которые необязательно имеют базовые навыки работы с компьютером и письма. Поэтому очень трудно обрабатывать полезную информацию.

Кроме того, все предложения требуют сначала построить модель в автономном режиме, а затем передать ее в онлайн. В реальном мире это не всегда возможно. Когда речь идет о достоверности и разнообразии данных, мы не всегда можем проанализировать автономные источники данных. Иногда приходится проводить анализ онлайн, что очень сложно, особенно учитывая перечисленные выше проблемы.

Заключение

В данной статье рассматриваются технические и интеллектуальные методы работы с данными разнородных типов. Комплекс разработанных методов интеллектуального анализа данных, используемый для анализа разнородных данных в рамках системы поиска и анализа данных, а также для улучшения процедур принятия решений для обработки большого потока данных с помощью методов и алгоритма, реализованных при разработке модуля, может улучшить процедуры принятия решений при обработке больших потоков данных. Эта система облегчит обработку больших наборов данных без необходимости использования нескольких различных программных пакетов для обеспечения однородности данных. Этапы обработки включают идентификацию данных в наборе данных, затем группировку их в классы в соответствии с типом, а затем, собственно, обработку. Для других текущих исследований она будет адаптирована к развёртыванию систем баз данных SQL server и NoSQL, которые предназначены для работы с информацией или данными с различными характеристиками.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Суварнамукхи Б. Концепции больших данных и методы обработки данных / Б. Суварнамукхи, М. Сешашаыее // International Journal of Computer Sciences and Engineering. - 2018. - Vol.6. - №10.

2. Ву Х. Интеллектуальный анализ данных с использованием больших данных / Х. Ву, Г. Чжу, У. Дин // IEEE Transactions on Knowledge and Data Engineering. – 2014.– №26 (1). – С. 97-107

3. Лейни Д. Управление 3d-данными: Управление объемом, скоростью и разнообразием данных / Д. Лейни // META Group Research Note 6. - 2001.

4. РБК Тренды: Что такое Большие данные и почему их называют «новой нефтью». -URL: https://trends.rbc.ru/trends/innovation/5d6c020b9a7947a740fea65c (дата обращения: 04.07.2015)

5. Qian Z.P. TimeStream: robust streaming computing in the cloud / Z.P. Qian , C.Z. Su et al. // the 8th ACM European Conference on Computer Systems. - 2013. - P. 1-4.

6. Чунг Д. Анализ больших данных: обзор литературы / Д. Чунг, Х. Ши // Journal of Management Analytics . – 2015. - Vol. 2. – № 3. – С. 175-201.

7. Романенко Е.В. Место Big Data в современной социально-экономической жизни общества / Е.В. Романенко // Инновационная наука. – 2016. – №4–3 (16). – С. 143–145.

8. Берман Дж. Проблемы конфиденциальности для сборщиков медицинских данных / Дж. Берман // Искусственный интеллект в медицине. – 2002. – №26. – С. 25-36.

9. Бабуцкий В. А. Методы и средства извлечения ключевых слов в задаче автоматической идентификации потенциально опасных текстов в условиях неопределенности их тематической принадлежности / В.А. Бабуцкий, И.Д. Сидоров // Успехи современной науки. – 2017. – Т. 1. – № 12. – С. 54–59.

10. Yang W. Discriminative topic model using document network structure / W. Yang , J. Boyd-Graber , P.A. Resnik // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. – 2016. – Vol. 1. – P. 686–696

11. Комарова А. В. Метод автоматизированного извлечения адресов из неструктурированных текстов /А. В. Комарова и др. // International Journal of Open Information Technologies. – 2017. – Т. 5. – № 11. – С. 21–26.

12. Piernik M. Clustering XML documents by patterns / M. Piernik, D. Brzezinski, T. Morzy // Knowledge and Information Systems. – 2016. – Vol. 46. – № 1. – P. 185–212.

13. Senta A. Python по байтам / A. Senta. - URL: https://pythobyte.com/python-script-to-organize-files-in-folders-5783-7c7c17a1/ (дата обращения: 27.12.2021)

14. Geeksforgeeks. - URL: https://www.geeksforgeeks.org/python-os-path-split-method/ (accessed: 21.10.2019)

15. eobermuhlner. - URL: https://github.com/eobermuhlner/kimage (accessed: 29.06. 2017)

16. ЖУРНАЛ OTUS. - URL: https://otus.ru/journal/instrumenty-python-dlya-raboty-s-izobrazheniyami/ (дата обращения: 11.04.2022)

17. Han J. Data mining: concepts and techniques / J. Han , J. Pei, M. Kamber. - Amsterdam : Elsevier, 2011.

18. Schlieski T. Entertainment in the age of Big Data / T. Schlieski, B.D. Johnson // Proceedings of the IEEE. - 2012. - Vol. 100. - P. 1404-1408

19. Гонсалес Р. Цифровая обработка изображений в среде MATLAB / Р. Гонсалес, Р. Вудс, С. Эддинс // Техносфера. - 2006. - 616 с.

Список литературы на английском языке / References in English

1. Suvarnamukhi B. Koncepcii bol'shih dannyh i metody obrabotki dannyh [Big data concepts and data processing methods] / B. Suvarnamukhi, M. Seshashayee // International Journal of Computer Sciences and Engineering. - 2018. - Vol.6.- №10. [in Russian]

2. Wu H. Intellektual'nyj analiz dannyh s ispol'zovaniem bol'shih dannyh [Data mining using big data] / H. Wu, G. Chzhu, U. Din // IEEE Transactions on Knowledge and Data Engineering. – 2014.– №26 (1). – P. 97-107 [in Russian]

3. Laney D. Upravlenie 3d-dannymi: Upravlenie ob'emom, skorost'ju i raznoobraziem dannyh [3d data management: Controlling data volume, velocity and variety] / D. Laney // // META Group Research Note 6. - 2001. [in Russian]

4. RBK Trendy: Chto takoe Bol'shie dannye i pochemu ih nazyvajut "novoj neft'ju" [What is Big Data and why they are called "new oil"]. -URL: https://trends.rbc.ru/trends/innovation/5d6c020b9a7947a740fea65c (accessed: 04.07.2015) [in Russian]

5. Qian Z.P. TimeStream: robust streaming computing in the cloud / Z.P. Qian , C.Z. Su et al. // the 8th ACM European Conference on Computer Systems. - 2013. - P. 1-4.

6. Chung D. Analiz bol'shih dannyh: obzor literatury [Big Data analysis: Literature review] / D. Chung, H. Shi // Journal of Management Analytics . – 2015. – Vol. 2. – № 3. – P. 175-201. [in Russian]

7. Romanenko E.V. Mesto Big Data v sovremennoj social'no-jekonomicheskoj zhizni obshhestva [The place of Big Data in the modern socio-economic life of society] / E.V. Romanenko // Innovacionnaja nauka[Innovative science]. – 2016. – №4–3 (16). – P. 143–145. [in Russian]

8. Berman Dzh. Problemy konfidencial'nosti dlja sborshhikov medicinskih dannyh [Privacy Concerns for Medical Data Collectors] / Dzh. Berman // Iskusstvennyj intellekt v medicine [Artificial intelligence in medicine]. – 2002. – №26. – P. 25-36. [in Russian]

9. Babuckij V. A. Metody i sredstva izvlechenija kljuchevyh slov v zadache avtomaticheskoj identifikacii potencial'no opasnyh tekstov v uslovijah neopredelennosti ih tematicheskoj prinadlezhnosti [Methods and means of extracting keywords in the task of automatic identification of potentially dangerous texts in conditions of uncertainty of their thematic affiliation] / V. A. Babuckij, I. D. Sidorov // Uspehi sovremennoj nauki[The successes of modern science]. – 2017. – Vol. 1. – № 12. – P. 54–59. [in Russian]

10. Yang W. Discriminative topic model using document network structure / W. Yang , J. Boyd-Graber , P.A. Resnik // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. – 2016. – Vol. 1. – P. 686–696

11. Komarova A. V. Metod avtomatizirovannogo izvlechenija adresov iz nestrukturirovannyh tekstov [Method of automated extraction of addresses from unstructured texts] /A. V. Komarova et al. // International Journal of Open Information Technologies. – 2017. – Vol. 5. – № 11. – P. 21–26. [in Russian]

12. Piernik M. Clustering XML documents by patterns / M. Piernik, D. Brzezinski, T. Morzy // Knowledge and Information Systems. – 2016. – Vol. 46. – № 1. – P. 185–212.

13. Senta A. Python po bajtam [Python by bytes] / A. Senta. -URL: https://pythobyte.com/python-script-to-organize-files-in-folders-5783-7c7c17a1/ (accessed: 27.12.2021) [in Russian]

14. Geeksforgeeks. - URL: https://www.geeksforgeeks.org/python-os-path-split-method/ (accessed: 21.10.2019)

15. eobermuhlner. - URL: https://github.com/eobermuhlner/kimage (accessed: 29.06. 2017)

16. ZhURNAL OTUS [Journal OTUS]. – URL: https://otus.ru/journal/instrumenty-python-dlya-raboty-s-izobrazheniyami/ (accessed: 11.04.2022) [in Russian]

17. Han J. Data mining: concepts and techniques / J. Han , J. Pei, M. Kamber. - Amsterdam : Elsevier, 2011.

18. Schlieski T. Entertainment in the age of Big Data / T. Schlieski, B.D. Johnson // Proceedings of the IEEE. - 2012. - Vol. 100. - P. 1404-1408

19. Gonsales R. Cifrovaja obrabotka izobrazhenij v srede MATLAB [Digital image processing in the MATLAB environment] / R. Gonsales , R.Vuds , S. Jeddins // Tehnosfera. - 2006. - 616 p. [in Russian]