Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет информатики и радиоэлектроники

УДК 004.62+004.438

Климец

Артём Александрович

АВТОМАТИЗАЦИЯ ПРОЦЕССОВ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ

АВТОРЕФЕРАТ

на соискание степени магистра

по специальности 1-40 80 01 Компьютерная инженерия. Хранение и обработка данных

Научный руководитель
Перцев Дмитрий Юрьевич
кандидат технических наук
доцент кафедры ЭВМ, БГУИР

Минск 2022

Нормоконтроль

Насуро Е.В.

ВВЕДЕНИЕ

Всемирная сеть Интернет стремительно вошла в жизнь современного общества и крепко укоренилась в ней. Каждый день сотни миллионов людей проводят значительную часть своей жизни в Интернете, генерируя огромное количество трафика. На данный счёт ежегодно проводятся исследования, прямо указывающие на общемировой тренд на всеобщую цифровизацию. Так, например, компания Domo представляет данные своих исследований в виде инфографики, отражающей, как много данных в различном виде генерируют пользователи различных сервисов. К примеру, в 2021 году пользователи социальной сети Facebook ежеминутно загружали в среднем 240 тысяч фотографий (в 2020 – 147 тыс.), так сумма транзакций на платформе Venmo увеличилась с 239 тыс. долларов до 304 тыс., а количество активных пользователей Microsoft Teams – с 52 до 100 тыс. ежеминутно [1][2].

Современный мир таков, что дальнейшее развитие человечества по большей части зависит от того, как будет использоваться вся имеющаяся информация, так как это огромный источник знаний, грамотное использование которых способствует получению определённой выгоды для тех, кто в этом заинтересован. Все создаваемые информационные потоки анализируются, обрабатываются и учитываются в различных сферах деятельности (реклама, прогноз успеха той или иной кампании и прочие). Так как «вручную» обработать весь поток информации невозможно, огромными темпами развивается такое направление, как интеллектуальный анализ данных, или, по-другому, Data Mining.

В данном направлении широкое развитие получили как коммерческие (например, MatLab, Statistica, RapidMiner), так и свободные (например, Weka, R, Torch и др.) специализированные инструменты, различные библиотеки и платформы. Однако они обладают как преимуществами в виду «заточки» на решение той или иной задачи, сопутствующая оптимизация связанных процессов, подготовленные модели, так и недостатками, к примеру, написание кода, усложнённый интерфейс со множеством параметров, работа только на персональном компьютере, скромные возможности для расширения функциональности. Данные инструменты не в состоянии полностью заменить разработку алгоритмов или построение собственных моделей, однако сильно упрощают работу как аналитикам, так и неквалифицированным работникам, которым в работе могут столкнуться с задачами по обработке данных.

Целью данной научной работы является изучение возможностей и подходов к автоматизации процессов интеллектуальной обработки данных. Для достижения данной цели необходимо решить следующие задачи: изучить процесс интеллектуального анализа данных и выявить этапы, которые следует ускорить, изучить существующие на рынке решения, выявить их сильные и слабые стороны. Результатом работы должен стать инструмент, который работал бы поверх какого-либо языка программирования и позволил бы упростить и ускорить процесс работы с данными. При выборе языка следует изучить аспекты применимости его для анализа данных и встраивания в исполняющийся код, для расширения базовой функциональности.

В качестве такого языка был выбран Python, как наиболее популярный среди специалистов в области анализа данных. Данный язык программирования обладает обширными методами интроспекции, позволяющими производить анализ самого языка и подключаемых к нему компонентов непосредственно во время исполнения программы.

Широкие возможности для расширения функциональности путём динамического подключения различных программных решений позволяют сделать инструмент по-настоящему универсальным, ведь добавление соответствующих библиотек языка программирования Python позволяет решать практически любую задачу. Простота и функциональность данной системы позволяет пользоваться ею людям, не обладающим соответствующими компетенциями в области работы с данными. К примеру, данной системой могли бы пользоваться студенты любых курсов, проводя какие-либо исследования, либо в рамках какой-либо учебной деятельности.

Рассмотрению в рамках данной работы подлежат следующие вопросы:

– Что подразумевает под собой интеллектуальный анализ данных?

– Из каких этапов состоит классический процесс интеллектуального анализа данных?

– Достаточно ли данных могут дать встроенные методы интроспекции языка Python?

– Может ли Python эффективно встраиваться в исполняемых код?

Ответы на данные вопросы послужат фундаментом предполагаемому инструменту.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью данной работы является исследование процесса интеллектуального анализа данных, областей применения и существующих инструментов для Data Mining. Задачей является разработка программного решения, особенностью и преимуществом над существующими решениями которого являются универсальность, возможность использования пакетов Python, размещённых в свободном каталоге PyPI в качестве основы для построения пользовательских алгоритмов обработки данных.

Исследование процесса интеллектуального анализа данных и возможностей его применения легло в основу доклада на научно-техническая конференция аспирантов, магистрантов и студентов БГУИР на тему «Современные методы анализа сетевого трафика», в котором рассмотрены возможности применения методов интеллектуального анализа данных в современных DPI-системах.

Разработанный в ходе научной работы инструмент предоставляет абстракцию над языком программирования Python, тем самым позволяя динамически подключать любую из существующих библиотек, на основе которых пользователь смог бы сформировать собственный алгоритм для обработки данных и отправить его на исполнение в динамически создаваемом окружении, позволяющее безопасно исполнять программный код на удалённых вычислительных ресурсах.

Данное решение имеет широкий спектр применений, но изначально подразумевается, что данный инструмент будет полезен специалистам по анализу данных, а также при осуществлении учебного процесса по предметам, связанным с наукой о данных.

Сведения о разработке данного инструмента были представлены на The 8th International Conference on ICT Management for Global Competitiveness and Economic Growth in Emerging Economies, прошедшей во Вроцлавском университете.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из 4 глав.

В первой главе рассматривается предметная область исследования: интеллектуальный анализ данных. В данной главе рассматриваются сферы применения, актуальность и то, какие задачи решаются методами интеллектуального анализа данных и как применяются различные инструменты для этого.

Вторая глава содержит в себе обзор и сравнительный анализ 14 различных инструментов для интеллектуального анализа данных, такие как WEKA, RapidMiner, KNIME, SAS Data Mining, H2O и другие. Описаны слабые и сильные стороны каждого из них, выделены технические решения, которые стоит воспроизвести, а также выделены возможности для улучшений, которые нашли своё отражение при создании собственного инструмента.

В третьей главе описаны основные идеи и концепции, которые легли в основу инструмента, названного Data Miner Combiner. Описаны структурные блоки и описаны варианты реализации для каждого блока, то есть описаны технологии или подходы, с помощью которых реализуется та или иная функциональность. Описаны вопросы производительности и безопасности исполнения стороннего кода.

В четвёртой главе описано то, каким инструмент получился в итоге, приведены скриншоты, иллюстрирующие работу платформы, а также описан этап функционального тестирования, подтверждающие соответствие системы заявленным требованиям, таким образом доказывая состоятельность концепции построения визуального инструмента для интеллектуального анализа данных в виде надстройки над существующим языком программирования.

ЗАКЛЮЧЕНИЕ

Исследование ситуации в среде программного обеспечения для проведения специализированных исследований в области Data Science показало, что на данный момент в сети Интернет нет достойных аналогов десктопных решений для проведения исследований, описанных в обзоре существующих аналогов разработанной системы. Поэтому реализация системы в виде веб-приложения является более чем логичным решением в современную эпоху развития сети Интернет, и повсеместного использования данной сети для самых разнообразных задач.

Результатом научной работы стал новый инструмент для интеллектуального анализа данных. Данная инструмент стал подтверждением гипотезы о том, что построение абстракции поверх языка программирования Python возможно и сам он может быть использован в качестве встраиваемого. Данное решение позволяет достичь большей универсальности и гибкости, по сравнению с большинством существующих решений.

Система позволяет упростить процесс проведения исследований с применением интеллектуального анализа данных, а также при проведении различного рода обработки информации, при этом в процессе интеллектуального анализа данных, работая на следующих этапах: подготовки данных, построения моделей и их применения на некотором массиве данных.

В качестве решения проблем с производительностью заложена возможность исполнять пользовательские алгоритмы на специализированном оборудовании, характеризующемся повышенными характеристиками аппаратного обеспечения, позволяя значительно ускорить процесс выполнения алгоритмов обработки данных, которые могут быть чрезвычайно ресурсоёмкими. Для решения проблем с безопасностью использована технология контейнеризации, позволяющая изолировать исполняющийся сторонний код, не нанося вред всей системе.

Дальнейшая работа над проектом будет заключаться в расширении возможностей выполнения вычислений на специализированных платформах. А также расширения пользовательских возможностей при работе с системой, в частности модернизация пользовательского интерфейса и построение полноценной среды визуального программирования.

Система может быть использована не только в рамках учебной и научной деятельности в университете и других учебных заведениях, но и в

коммерческих целях после незначительной доработки, главным образом связанной с безопасностью использования системой.

Главным недостатком инструмента на данный момент является его низкое быстродействие, однако данный недостаток сложно в полной мере компенсировать, так как основой построения всей системы в целом является использование библиотек Python, а данный язык программирования не отличается своим быстродействием. Помимо это существуют задержки при запуске алгоритмов в контейнерах Docker, ведь несмотря на свою легковесность, контейнеры всё же требуют некоторое время для запуска и инициализации, прежде чем они будут готовы к работе.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

[1-А.] Климец, А. А. Современные методы анализа сетевого трафика / А.А. Климец // 57-я конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», 19-23 апреля 2021 г., БГУИР, Минск, Беларусь: тез. докл. – Мн. – 2021. – 155с.

[2-А.] Klimets, A. Data Mining as a Service / A. Klimets // The digital innovation in information systems, virtual management and their impact on sustainable development in crisis conditions: proceedings of the 8th International Conference on ICT Management for Global Competitiveness and Economic Growth in Emerging Economies ICTM 2021, Wroclaw, 26-28 Oct. 2021. – Wroclaw: University of Wroclaw, 2021. – P. 82.