УДК

Перова Марина Викторовна
Заведующая кафедры информационных технологий
Южно-Российский институт управления – филиал Российской академии народного хозяйства и государственной службы при Президенте РФ, Ростов-на-Дону
Email: _____

Жуковский Сергей Романович
Студент
Южно-Российский институт управления – филиал Российской академии народного хозяйства и государственной службы при Президенте РФ, Ростов-на-Дону
Email: _____

OPTICAL CHARACTER RECOGNITION (OCR) РЕШЕНИЯ И ИХ ИСПОЛЬЗОВАНИЕ В СЭД

Аннотация: В данной статье анализируется технология (модуль) OCR (ОРС) как инструмент оптимизации работы с большим потоком документов и технологии автоматизации управленческих и организационных процессов государственного сектора и бизнес-структур. Технология оптического распознавания символов имеет определенные этапы считывания текстовой информации и технологии, которые оптимизируют данные этапы, минимизируя риск ошибок распознавания информации и, как следствие, оцифровки документов. Использование модуля ОРС совместно с технологиями машинного обучения, нейросетями, модулем ИИ для СЭД и других новшеств современного ИТ-рынка позволяет реализовать масштабирование и интеграцию ОРС в различных сценариях использования, а также увеличить количество сценариев использования модуля, тем самым позволив органам власти и бизнесу высвободить значительное количество временных, финансовых и кадровых ресурсов.

Ключевые слова: OCR, нейросети, машинное обучение, система электронного документооборота, распознавание информации, API.

OPTICAL CHARACTER RECOGNITION (OCR) SOLUTIONS AND THEIR USE IN EDMS

Annotation: This article analyzes the technology (module) OCR (OCR) as a tool for optimizing work with a large flow of documents and technology for automating

managerial and organizational processes in the public sector and business structures. Optical character recognition technology has certain stages of reading text information and technologies that optimize these stages, minimizing the risk of information recognition errors and, as a result, digitization of documents. The use of the LFS module in conjunction with machine learning technologies, neural networks, the AI module for EDMS and other innovations of the modern IT market allows for scaling and integration of the LFS in various use scenarios, as well as increasing the number of scenarios for using the module, thereby allowing authorities and businesses to free up a significant amount of time, financial and human resources.

Keywords: OCR, neural networks, machine learning, electronic document management system, information recognition, API.

В современном мире использование систем электронного документооборота (СЭД), а также различных технологий обработки электронной информации становится все более распространено за счет политики импортозамещения на рынке ИТ-решений в РФ, а также желанием государственного сектора и бизнеса оптимизировать управленческие и организационные процессы, тем самым высвободив значительное количество кадровых, временных и финансовых ресурсов.

При этом вопросы, связанные с документационным обеспечением органов власти и бизнеса, касающиеся оцифровки текстовых файлов, поступающей документации, автономного распознавания текста и перевода его в удобный для пользователя цифровой формат имеют огромное значение в рамках процесса обработки документации. Современное ИТ-решение optical character recognition – OCR (Оптическое распознавание символов - ОРС) – выступая модулем СЭД или микросервисной технологией значительно упрощает входящую обработку документации и последующую работу с ней. Технология заключается в интеллектуальном распознавании не оцифрованного документа посредством предоставления API для точной идентификации символов с изображений и отсканированных документов на различных языках и форматах [1, с. 6]. Проще говоря интеллектуальная система при распознавании текста в документах сканирует текстовый слой посимвольно и вытаскивает его из документа для дальнейшего редактирования. В случае с изображениями (сканами документов), интеллектуальная система разбивает его на блоки с помощью алгоритма

топологического структурного анализа бинарных изображений [2, с. 46]. Далее на каждом блоке выделенного изображения происходит распознавание текстовых символов с помощью системной библиотеки.

В соответствии с Указом Президента от 10.10.2019 года «О развитии искусственного интеллекта в Российской Федерации», а также Национальным проектом «Цифровая экономика» развитие цифровых технологий и искусственного интеллекта становятся приоритетными направлениями развития экономики и государственного управления. Это касается прежде всего нейросетевого взаимодействия, технологий искусственного интеллекта, машинного обучения, цифровых инфраструктурных решений для бизнеса и государства и многого другого. Использование новых цифровых технологий в российском государственном секторе и экономических отношениях выходят на передовые позиции государственной политики, как средства технологического совершенствования [7].

Российский рынок OCR-решений не стал исключением и был вовлечен в развитии ИТ-рынка РФ в целом и отдельных компаний после ухода из РФ крупномасштабного разработчика серии продуктов FineReader – компании ABBYY. На 2022 год выручка разработчиков OCR-продуктов упала на 24% по сравнению с 2021 годом, так как компания ABBYY занимала более 2/3 всего рынка соответствующих технологий [6]. Однако, активная политика импортозамещения и стремление компаний к независимости от зарубежных продуктов привели в 2023 году к взлету рынка финтех-разработчиков и вендоров на 43% по сравнению с 2022 годом (рис. 1).

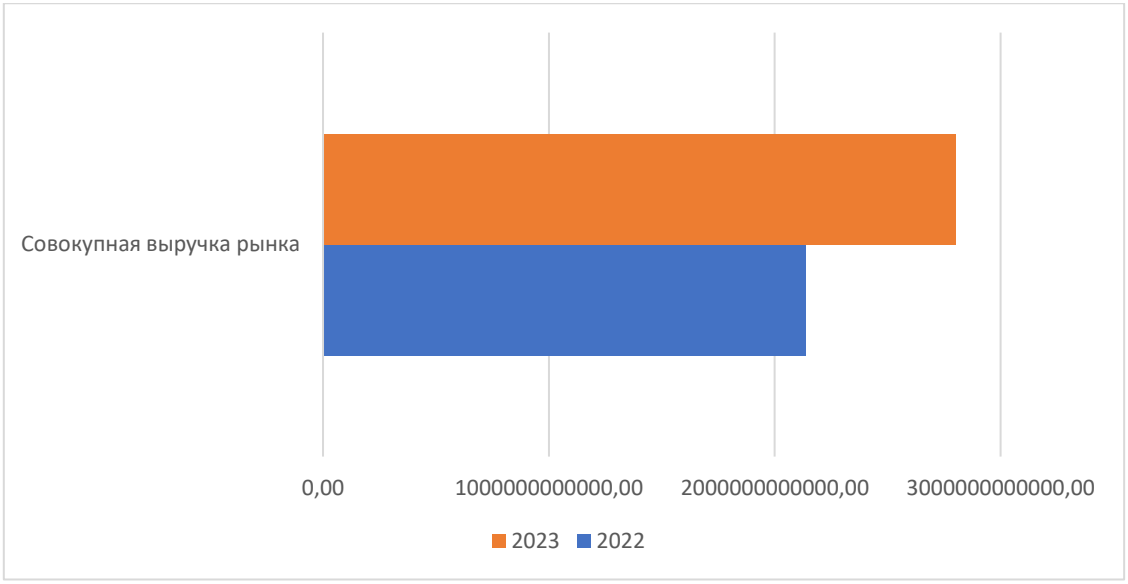

Рис. 1. Динамика совокупной выручки ИТ-рынка в РФ [8].

Современные OCR-решения продукты используются повсеместно и в государственном секторе, экономике, бизнесе. Один из флагманских продуктов Content AI — ContentCapture, универсальная платформа для интеллектуальной обработки информации. Решение, призванное заменить продукт ABBYY FlexiCapture, включено в дорожную карту развития и поддержки нового общесистемного программного обеспечения (НОПО).Продукт Content AI отнесен к разряду общесистемного ПО, поскольку он комплексно решает целый ряд задач, связанных с обработкой бухгалтерских, финансовых, юридических документов как для крупного, так и для среднего бизнеса.ContentCapture обрабатывает в едином потоке бумажные и электронные документы любых типов: отсканированные бумаги, фотографии, тексты, PDF-файлы, полученные из разных источников — электронных писем, FTP, «горячих папок», мобильных устройств, МФУ, сетевых сканеров. Причем разные форматы документов можно обрабатывать одновременно [9].

В 2023 году команда СЭД Тезис активно вела разработки в области искусственного интеллекта и его использования в СЭД и цифровой экономике. Встроенные модули ИИ, одним из которых является использование OCR совместно с ИТ-инфраструктурой компании или органа власти, позволяют оптимизировать значительную часть процессов по управлению и хранению данными, потоками документов, а также решение текущих процессов на основе

обучения систем искусственного интеллекта в конкретной системе. Не исключением становится OCR-модуль, который основан на технологиях искусственного интеллекта и являющийся одним из передовых направлений разработки программного обеспечения и его комплектующих для бизнеса и государственного управления. На 2024 год модули ИИ внесены в Роспатент и используются в рамках государственных программ по цифровизации государственного и муниципального управления, в цифровой экономике, в бизнес-процессах, а также во взаимодействии общества и государства. Как раз таки развитие технологий искусственного интеллекта в РФ усилило необходимость и желание российских компаний привлечь дополнительные инструменты для оптимизации процессов, в том числе оцифровке бумажной документации [10].

ОРС-решение в качестве модуля СЭД имеет собственный алгоритм оптимизации и может быть связан с машинным обучением, искусственным интеллектом (ИИ) и интегрированными системами электронного документооборота (ЭДО) для автоматизации не только распознавания выходных документов, но и работы с архивной документацией, обращениями и внутриведомственной текущей документацией.

ОРС модуль имеет 3 стадии обработки изображения или документа:

1. Обработка изображений. В основном при обработке изображений используются 2 технологии: функциональное выравнивание Гаусса и бинаризация входящего изображения. Первая технология позволяет избавиться от неровного расположения текстовых полей отсканированного документа и минимизировать «шум» для более четкой оцифровки. Вторая же направлена на считывание символов текста и перевод их из цветного формата в черно-белый для упрощения последующей работы с текстом.

2. Сегментация. На втором этапе изображение проходит компьютерно-математическую обработку на основании проверки корректности текста и его основных параметров. То есть изображение делится на бинарные топологические единицы и с помощью модуля ИИ и математических

вычислений данных изображения производится обработка корректности текстовых строк и параметров документа.

3. Распознавание символов. В основном распознавание символов основаны на методе интеллектуального извлечения, где каждый символ считывается и извлекается на основе определенных систематических правил семантики и графической структуры изображения – пиксельной схемы сканированного документа [3, с. 90].

Сама технология ОРС имеет несколько сценариев использования, популярными из которых являются: оцифровка архивов, классификация и маршрутизация текущих документов, полнотекстовая сверка и сверка с шаблоном, обработка первичной документации, обработка кадровых документов, выверка данных и модификация документов. Данные сценарии представлены командной SOICA компании SL Soft [4]. По их мнению, использование ОРС решений в данных направлениях позволяет автоматизировать государственные и бизнес управленческие процессы, а также интегрировать новую среду ИИ в СЭД. Сама же технология ОРС на базе каст-программного обеспечения отечественных разработчиков способна к развертыванию и масштабированию за счет крупной контейнеризации и, в случае необходимости, технология может быть перемещена на новые мощности для увеличения пропускной способности или же при создании нового контейнера модуля ОРС.

Примером масштабирования технологии является интеграция решения российского вендора Content AI совместно с крупнейшим в РФ поставщиком ИТ-решений ЭОС. Компания ЭОС официально перешла на использование в своих решениях российского продукта ContentReader Engine — многофункционального OCR SDK для встраивания в приложения функций интеллектуального распознавания информации. Технологии ContentReader Engine применяются в модуле потокового ввода данных в СЭД «ДЕЛО». С их помощью пользователи СЭД могут быстро и точно распознавать информацию в разных типах документов, изображений, фотографий, скриншотов, мониторов и

дисплеев с сохранением в форматах Word, Excel, PDF и других. Это позволяет в дальнейшем осуществлять полнотекстовый поиск по документам [5].

Таким образом, использование OCR модуля является важным компонентом систем электронного документооборота (СЭД), так как позволяет автоматически распознавать текст из отсканированных документов. Это упрощает процесс работы с большим объемом документов, так как устраняется необходимость вручную перепечатывать информацию. Однако, при выборе OCR решения для СЭД следует учитывать несколько важных факторов. Важно, чтобы система обладала высокой точностью распознавания текста, а также поддерживала различные языки и форматы документов. Также стоит обратить внимание на скорость работы и интеграционные возможности с другими компонентами СЭД, что позволит внедрять и масштабировать данную технологию в различные сценарии использования.

Список используемой литературы

1. Аверьянова А.Н., Атанов В.В., Кеся М.С., Можнов Е.С. Использование интеллектуальных микросервисов в современных системах электронного документооборота // Форум молодых ученых. 2023. №5. URL: https://cyberleninka.ru/article/n/ispolzovanie-intellektualnyh-mikroservisov-v-sovremennyh-sistemah-elektronnogo-dokumentooborota (дата обращения: 25.04.2024).

2. Бикметова М.Р., Азбуханов А.Ф. Использование машинного обучения для автоматизации электронного документооборота на предприятии // Информационные технологии: проблемы и решения. Уфа. 2021. №3. С. 76

3. Хоменко Т.В., Иргалиев А. А., Тараканов В.в Д. Моделирование процесса распознавания символов в нормативных документах организации // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2023. №2. URL: https://cyberleninka.ru/article/n/modelirovanie-protsessa-raspoznavaniya-simvolov-v-normativnyh-dokumentah-organizatsii (дата обращения: 25.04.2024).

4. SOICA: как OCR помогает автоматизировать документооборот. Tadviser. Официальный сайт – [Электронный ресурс] – Режим доступа: www.tadviser.ru (Дата обращения: 25.04.2024)

5. Российский вендор Content AI будет внедрять свои технологии в СЭД/ECM-решения разработчика систем управления документооборотом ЭОС. ЭОС. Официальный сайт. – [Электронный ресурс] – Режим доступа: eos.ru (Дата обращения: 25.04.2024)

6. Россия без FineReader: рынок OCR за год упал на четверть. Cnews. Официальный сайт. – [Электронный ресурс] – Режим доступа: www.cnews.ru (Дата обращения: 25.04.2024)

7. Указ Президента РФ от 10.10.2019. «О развитии искусственного интеллекта (совместно с Национальной стратегией развития искусственного интеллекта на период до 2030 года)». // Информационно-правовой портал Гарант

8. Выручка ИТ-компаний в России выросла на 43%. Минцифры. Официальный портал. – [Электронный ресурс] – Режим доступа: https://digital.gov.ru/ (дата обращения: 28.05.2024)

9. ContentCapture — универсальный инструмент для интеллектуальной обработки документов. Tadviser. Официальный сайт – [Электронный ресурс] – Режим доступа: www.tadviser.ru (Дата обращения: 28.05.2024)

10. СЭД ТЕЗИС: Возможности модуля искусственного интеллекта. Tadviser. Официальный сайт – [Электронный ресурс] – Режим доступа: www.tadviser.ru (Дата обращения: 30.05.2024)