Ключевой оказывается проверка условия $|M| = N - 1$. Задача сведения системы (1) к системе (4) разрешима [1,2]:

- стохастически однозначно при $N < |M| + 1$;
- детерминированно однозначно при $N = |M| + 1$;
- не разрешима при $N > |M| + 1$.

Соотношения (5) позволяют ставить обратную задачу подбора латентных экзогенных регрессоров по характеристикам наблюдаемых эндогенных переменных [3,4]. Достаточно лишь подобрать необходимое число $|M|$ полиномиальных слагаемых и воспользоваться уравнениями (2), (3) для включения отобранных степеней в систему (1).

Литература

1. Котенко А.П., Букаренко М.Б. Геометрия систем линейных регрессионных уравнений / А.П. Котенко, М.Б. Букаренко // Известия Самарского научного центра РАН. – 2013. – т.3, №6(3). – С.820-823.

2. Котенко А.П. Особенности применения косвенного метода наименьших квадратов к системе независимых эконометрических уравнений / А.П. Котенко // Друкеровский вестник. – 2017. – №3. – С.96-102.

3. Котенко А.П., Кузнецова О.А. Применение методов многомерного регрессионного анализа для оптимизации производства битума стандартизованных характеристик / А.П. Котенко, О.А. Кузнецова // Современные информационные технологии и ИТ-образование. Сб. научных трудов. – М.: Изд-во МГУ. – 2015. – С.381-384.

4. Котенко А.П., Котенко А.А. Использование идентифицируемых систем эконометрических уравнений / А.П. Котенко, А.А. Котенко // Математика, статистика и информационные технологии в экономике, управлении и образовании. Сб. трудов V Международной научно-практической конф. – Тверь: Изд-во Тверского гос. ун-та, 2016. – С.51-55.

Д.М. Кусаинов, А.А. Столбова

АВТОМАТИЗИРОВАННАЯ СИСТЕМА РАСПОЗНАВАНИЯ ТЕКСТА С ТАБЛИЧНЫМИ СТРУКТУРАМИ НА ИЗОБРАЖЕНИЯХ

(Самарский университет)

Во многих сферах деятельности происходит стремительный рост информационных потоков, что влечет за собой необходимость создания систем электронного документооборота [1]. Кроме того, цифровизация является повсеместной тенденцией, что подтверждается развитием таких подходов как «Индустрия 4.0» и «Интернет вещей» [2]. Во многих производственных компаниях и не только распространены бумажные варианты документов, которые требуют перевода в электронный вид. В таких документах данные

могут быть представлены различными способами: текст, изображения, таблицы. При оцифровке документов важным является не только получение цифровой копии документа, но и оптическое распознавание текста с извлечением данных из него для последующей обработки. Одной из сложных задач в данной области, требующих решения, является распознавание табличных структур.

Существуют различные автоматизированные системы, позволяющие распознавать документы, представленные в виде изображения (например, скан документа). Одной из таких систем является «ABBYY FineReader», отличающаяся своим быстродействием и точностью распознавания. Система делит текст на строки, слова и символы, после чего включаются механизмы распознавания – классификаторы [3]. Недостатком данной системы является ее дороговизна. Другим примером подобных систем является отечественная разработка для оптического распознавания текста – «OCR CuneiForm». Программа поддерживает работу с таблицами, а именно, автоматически находит их в тексте и распознает, позволяет работать с изображениями в исходных документах, а также имеет встроенный редактор таблиц [4]. Данная система представляет собой свободно распространяемое программное обеспечение, однако не поддерживается операционной системой MS Windows.

В рамках данной работы предлагается разработка автоматизированной системы распознавания текста с табличными структурами на изображениях.

Разрабатываемая система поддерживает работу через систему личных кабинетов с поддержкой следующих функциональных возможностей:

– работа с исходными документами в форматах pdf, jpg: загрузка, удаление;
– распознавание документов в форматах pdf, jpg;
– редактирование текста распознанного документа;
– сохранение результатов распознавания в формате txt.

Автоматизированная система распознавания текста с табличными структурами на изображениях представляет собой веб-приложение. На рисунке 1 представлен прототип главной экранной формы системы. Разработка ведется в интегрированной среде разработки программного обеспечения Intellij IDEA на объектно-ориентированном языке программирования Java с использованием технологий AngularCLI и Hibernate.

Процесс распознавания документов включает в себя следующие этапы:

– импорт исходного документа в заданном формате;
– определение структуры документа: наличие таблиц, абзацев;
– распознавание документа;
– проверка и редактирование результатов распознавания;
– экспорт полученного результата.

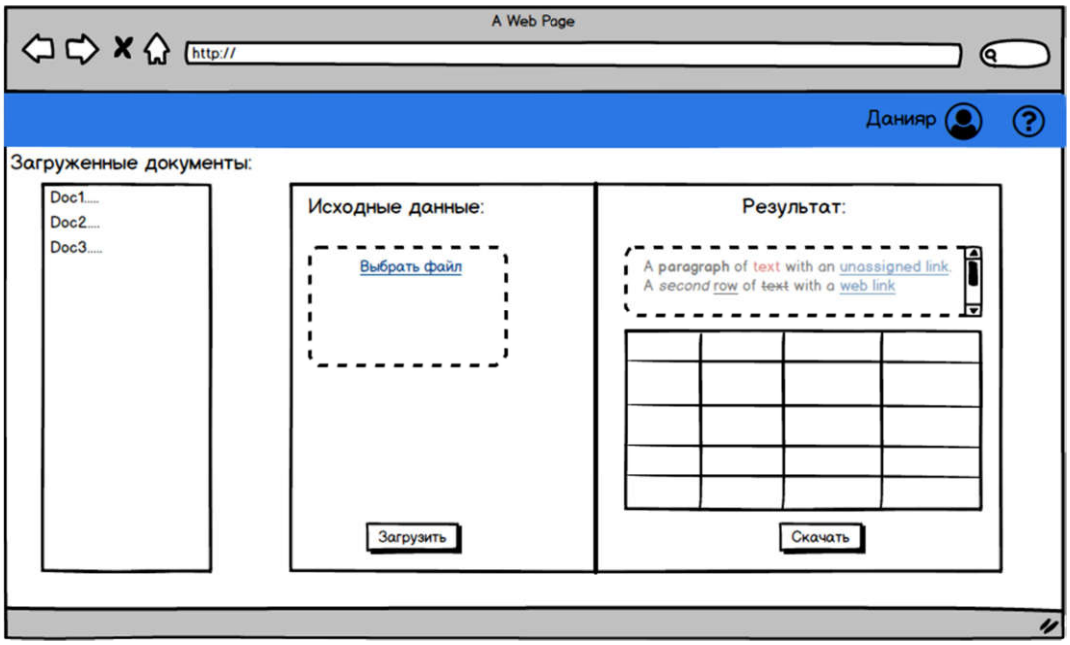

Рис. 1. Прототип экранной формы системы

Таким образом, в рамках проведенной работы предложен проект автоматизированной системы распознавания текста с табличными структурами на изображениях, рассмотрены этапы оптического распознавания документов, содержащих табличные структуры, произведен анализ аналогичных систем.

Литература

10 Клименков С. В., Ткешелашвили Н. М., Дергачев А. М. Метод распознавания структуры таблицы в электронных табличных документах //Программные продукты и системы. – 2016. – №. 4 (116).].

11 Андиева Е. Ю., Фильчакова В. Д. Цифровая экономика будущего, индустрия 4.0 //Прикладная математика и фундаментальная информатика. – 2016. – №. 3. – С. 214-218

12 Распознавание текста с помощью решений ABYY – всё гениально просто для бизнеса [Электронный ресурс] URL: https://www.kp.ru/guide/raspoznavanie-teksta.html (дата обращения: 08.04.2020).

13 Распознавание текста – OCR CunneiForm [Электронный ресурс] // URL: http://pro-spo.ru/text/341--ocr-cuneiform (дата обращения: 08.04.2020).

14 Программы распознавания текста [Электронный ресурс] // URL: https://www.it-world.ru/tech/admin/139030.html (дата обращения: 08.04.2020).