Отзыв официального оппонента

на диссертационную работу Беляевой Оксаны Владимировны

«Автоматическое восстановление структуры текстовых документов», представленную на соискание ученой степени кандидата технических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Актуальность темы диссертации

Современный этап развития информационных технологий характеризуется постоянным ростом объемов неструктурированных данных, значительная часть которых представлена в виде печатно-ориентированных документов (в том числе в формате PDF). Несмотря на то, что PDF является де-факто стандартом электронного документооборота, его внутренняя специфика (ориентированность на визуальное отображение, а не на хранение логической структуры) создает серьезные барьеры для автоматизированной обработки данных. PDF-документы хранят информацию в виде низкоуровневых инструкций отрисовки текста и графики (часто без аннотирования логической структуры — иерархии заголовков, абзацев, таблиц и пр.). В то же время современные приложения, в которых такие данные находят применение, часто требуют представить их в полуструктурированном или структурированном виде. Автоматическое восстановление логической структуры таких документов является критически важной задачей процессов преобразования неструктурированной документной информации к требуемой форме.

Существующие технологические решения зачастую демонстрируют недостаточную эффективность (по затрачиваемым вычислительным ресурсам) и надежность (по диапазону обрабатываемых случаев) при обработке PDF-документов с разнородным содержимым (растровыми изображениями страниц, текстовым слоем, встроенными кодировками символов и пр.). В основном существующие технологические решения следуют одному из двух подходов: одни конвертируют PDF-документы в растр и затем применяют оптическое распознавание символов (OCR); другие извлекают содержимое напрямую из PDF с помощью выполнения инструкций отрисовки текста и графики. С одной стороны, решения на основе OCR менее эффективны по сравнению с решениями на основе прямого доступа, с другой стороны, решения на основе прямого доступа к PDF менее надежны по сравнению с решениями на основе OCR. Таким образом, разработка надежных и эффективных (в том числе, гибридных) методов и инструментальных средств автоматического извлечения содержимого и восстановления логической структуры печатно-ориентированных документов PDF является актуальной научно-технической задачей, решение которой позволит существенно повысить эффективность и надежность автоматизированной обработки документной информации.

Обоснованность и достоверность научных результатов диссертации

В диссертации представлены следующие научные результаты: метод автоматического извлечения содержимого PDF-документов с проверкой корректности текстового слоя; метод восстановления иерархической структуры текстовых документов; архитектура и расширяемый программный комплекс для автоматической обработки документов различных форматов и предметных областей. Обоснованность научных результатов подтверждается следующим. Для метода извлечения содержимого проведено экспериментальное

сравнение трех режимов обработки на тестовых наборах документов: предложенный гибридный метод обеспечивает точность извлечения текста (character accuracy) 0.939 против 0.906 в режиме OCR и 0.589 в режиме прямого чтения PDF (без OCR) на «двухстраничном» наборе данных, при этом скорость обработки возрастает более чем в 5 раз (по сравнению с режимом OCR) на наборе данных реальных документов. Сравнение с открытыми системами на некорректных PDF показало character accuracy 0.91 у разработанного комплекса против 0.08–0.097 у конкурентных решений. Метод восстановления иерархической структуры оценен на размеченных наборах трех предметных областей (законы РФ, технические задания, дипломные работы) с точностью классификации строк от 0.88 до 0.95. На наборе данных соревнования FINTOC 2022 предложенный метод превзошел всех участников: F1-мера обнаружения заголовков составила 0.900 против 0.830 у ближайшего конкурента, а точность определения уровня заголовка (level accuracy) — 58.4 против 42.9 у ближайшего конкурента. Достоверность результатов подтверждается внедрением в четырёх организациях, интеграцией в открытую библиотеку LangChain, апробацией на девяти конференциях (2019–2024 гг.) и публикацией в 5 журналах, рекомендованных ВАК, из которых 3 индексируются в Scopus и Web of Science.

Новизна научных результатов диссертации

Новизна научных результатов диссертационной работы заключается в комплексном решении научной задачи автоматического анализа структуры печатно-ориентированных документов с учетом настраиваемой специфики разных форматов и предметных областей. К новым научным результатам, полученным автором диссертации, следует отнести следующие:

1. Предложен новый метод автоматического извлечения содержимого PDF-документов с учетом разнородности их представления. В отличие от известных подходов, впервые выбор наиболее эффективного и надежного режима обработки PDF-документа осуществляется автоматически на основе бинарного классификатора в зависимости от исходного представления: растровое или текстовое представление, встроенные или стандартные кодировки символов, наличие искажений текста и пр. В результате обеспечивается более эффективная по времени выполнения обработка PDF-документа по сравнению с базовыми решениями на основе OCR и более надежная по поддерживаемым входным данным обработка PDF-документа по сравнению с базовыми решениями на основе прямого обращения к инструкциям отрисовки PDF-документа.
2. Предложен новый метод автоматического восстановления иерархической структуры содержимого печатно-ориентированных документов с учетом их предметной специфики. Предложенный метод конкретизируется рядом методик классификации строк на основе компоновочных и стилевых признаков, а также регулярных выражений, характерных для выбранной предметной области. Технологическое решение, реализующее предлагаемый метод, показывает высокое качество восстановления структуры, по сравнению с имеющимися аналогами, на наборе данных международного соревнования FINTOC2022.

Замечания по диссертации

1. Предлагаемый метод извлечения содержимого PDF-документа сначала оценивает «корректность» (т.е. доступность для обработки без OCR) всего документа в целом, а затем принимает решение, каким способом обрабатывать все страницы документа кроме первой: (1) через растеризацию PDF (с OCR) или (2) напрямую через инструкции отрисовки PDF (без OCR). Однако, в общем случае, внутри PDF-документа

(более того внутри одной страницы) может быть представлено разнородное содержимое: текст, вставленный в виде растра, и текст, выводимый инструкциями отрисовки (со встроенными CMAP и стандартными кодировками). Вероятно, что в некоторых случаях предпочтительнее было бы оценивать «корректность» и соответственно выбирать способ обработки каждой страницы в отдельности, а не всего документа в целом. В работе следовало бы рассмотреть ограничения предлагаемого метода в явной форме, в особенности те случаи, которые могут повлиять на эффективность и надежность обработки документов.

2. В диссертации приводятся сведения о практическом применении полученных автором диссертации научных результатов только в виде их краткого перечисления и копий подтверждений их внедрения. Поскольку представленные результаты имеют не только теоретический, но и прикладной характер, в диссертационной работе было бы уместно рассмотреть случаи их практического применения более подробно.
3. В работе указано, что «объектом исследования являются текстовые электронные документы различных предметных областей в виде структурированных и неструктурированных форматов документов». Судя по представленной работе, объектом исследования являются скорее процессы анализа структуры таких документов (автоматизации именно этих процессов посвящена диссертационная работа).
4. В формулировке первого из основных результатов и его научной новизны указано, что предложенный метод обеспечивает «достоверность извлечения и скорость обработки» без каких-либо качественных определений достигаемых эффектов, например: «обеспечивается *более полная* достоверность извлечения по сравнению с аналогами» или «обеспечивается *более высокая* скорость обработки по сравнению с аналогами». (Количественные определения этих эффектов приведены в тексте диссертации.)
5. В работе иногда используются жаргонизмы вместо словарных терминов русского языка, например: «параграф» вместо «абзац», «детекция/детектирование» вместо «обнаружение», «кракозябры» вместо «некорректно отображаемых символов», «сервер поднимается» вместо «сервер запускается».

Указанные замечания не влияют на общую положительную оценку данной работы.

Заключение по диссертации

Представленная диссертация является научно-квалификационной работой, в которой содержится решение научной задачи автоматического анализа структуры печатно-ориентированных документов с учетом настраиваемой специфики разных форматов и предметных областей, имеющей важное значение для развития знаний в области математического и программного обеспечения вычислительных систем, комплексов и компьютерных сетей. Более того, в диссертации изложено новое научно обоснованное технологическое решение, а именно расширяемый программный комплекс для автоматического извлечения содержимого документов и восстановления их структуры, обеспечивающее повышение эффективности и надежности процессов обработки документной информации разных форматов и предметных областей. Предложенные методы и разработанный на их основе программный комплекс имеют существенное значение для развития страны в сфере автоматизации обработки архивных документов, технических регламентов, научной периодики, нормативно-правовых актов и прочих видов документной информации в рамках программ импортозамещения программного обеспечения и построения суверенных баз данных и знаний.

Диссертация обладает внутренним единством, все представленные результаты согласованы с поставленной целью и задачами исследования. Автором диссертации выдвигаются для защиты новые научные результаты, которые соответствуют направлениям исследований паспорта специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Диссертация свидетельствует о личном вкладе ее автора в науку по направлениям данной специальности.

Предложенные автором решения автоматизации процессов извлечения содержимого и восстановления структуры печатно-ориентированных документов аргументированы представленными в диссертации оценками производительности и сравнениями с известными аналогами. В диссертации приводятся сведения о практическом использовании полученных автором диссертации научных результатов. Программное обеспечение, разработанное в рамках диссертационного исследования, внедрено в четырех цифровых продуктах. Соответствующие копии подтверждений внедрения представлены в приложении к диссертации.

Основные научные результаты диссертации опубликованы в пяти рецензируемых научных изданиях, рекомендованных ВАК. Программное обеспечение, разработанное в рамках диссертационного исследования, зарегистрировано в Роспатенте: получено три свидетельства о государственной регистрации программы для ЭВМ. В диссертации приводится описание личного вклада автора и вклада его соавторов в результаты совместных научных работ. В диссертации автор корректно ссылается на источники использованных им материалов.

Диссертационная работа соответствует требованиям Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842 (в действующей редакции), предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а ее автор, Беляева Оксана Владимировна, заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент:

Шигаров Алексей Олегович,

Кандидат технических наук, ведущий научный сотрудник лаборатории комплексных информационных систем Института динамики систем и теории управления имени В. М. Матросова Сибирского отделения Российской академии наук (ИДСТУ СО РАН).

Контактная информация:

Федеральное государственное бюджетное учреждение науки Институт динамики систем и теории управления имени В. М. Матросова Сибирского отделения Российской академии наук, 664033, г. Иркутск, ул. Лермонтова, 134. Телефон: +7 (3952) 42-71-00. Эл. почта: idstu@icc.ru.

А. О. Шигаров

«13» марта 2026 г.

Подпись заверяю
Нач. отдела делопроизводства
и организационного обеспечения

Г.Б. Кононенко
13.03.2026