ОТЗЫВ

официального оппонента на диссертацию

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. С ростом объема хранимых неструктурированных данных задачи автоматической обработки электронных текстовых документов становятся особенно востребованными. Это подтверждается активным развитием исследований в данной области, которые сосредоточены на создании эффективных методов анализа, извлечения и структурирования текстовой информации, содержащихся в электронных документах. Результаты этих исследований широко применяются в системах интеллектуальной обработки текстов, что позволяет значительно повысить их качество и скорость работы.

Особую важность представляет не только извлечение информации, но и восстановление иерархической структуры документов, что является ключевым аспектом для систем анализа больших данных. Для эффективного выполнения задач быстрого поиска и обработки актуальной информации такие системы требуют иерархического фрагментирования содержимого документов, что обеспечивает упорядоченность данных и их удобство для дальнейшего анализа.

В этом контексте работа Оксаны Владимировны Беляевой, посвященная разработке методов автоматического извлечения содержимого и восстановления структуры текстовых документов, безусловно является актуальной. Проведенные в диссертации исследования отвечают на вызовы, связанные со стремительным ростом объема электронных документов, и

предлагают решения, позволяющие значительно упростить и ускорить процессы обработки и анализа данных.

Научная новизна. В диссертационной работе получены следующие наиболее важные научные результаты.

1. Разработан новый метод автоматического извлечения содержимого PDF-документов с использованием проверки текстового слоя, обеспечивающий повышение достоверности извлечения и высокую скорость обработки документов на русском и английском языках.

2. Предложен новый метод автоматического восстановления иерархической структуры из содержимого документов. Метод показывает более высокое качество восстановления структуры по сравнению с другими методами, что подтверждается результатами экспериментов, в том числе, на наборе данных международного соревнования FinTOC-2022.

Практическая значимость. В ходе диссертационного исследования автором разработано программное средство автоматического извлечения содержимого и восстановления иерархической структуры текстовых документов. Разработанное средство внедрено в нескольких организациях, что подтверждается актами о внедрении.

Достоверность и обоснованность научных результатов. В диссертационной работе корректно применяется математический аппарат, лежащий в основе используемых нейросетевых моделей, методов машинного обучения и средств автоматической обработки документов. Экспериментальные исследования разработанных методов проводятся на нескольких тестовых наборах данных с применением адекватной методологии. Полученные результаты экспериментов сопоставлены с известными лучшими результатами и показано превосходство предложенных в диссертации решений.

Основные результаты диссертации обсуждались на научных конференциях, в том числе международных. Содержание исследования достаточно полно изложено в десяти печатных изданиях, в том числе опубликовано пять статей в изданиях из перечня ВАК. Имеется 3 свидетельства о государственной регистрации программ для ЭВМ.

Содержание работы.

Диссертация включает введение, 3 главы, заключение, библиографический список и приложения.

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследования, показаны научная новизна и практическая значимость работы.

В первой главе представлен обзор существующих методов извлечения содержимого из документов PDF-формата и изображений, а также методов восстановления иерархической структуры документа. Описаны систем автоматической обработки документов и выполнено их сравнение.

Вторая глава посвящена разработке методов автоматической обработки электронных текстовых документов. Представлен метод автоматического извлечения содержимого PDF-документов с использованием проверки текстового слоя. Предложен метод восстановления иерархической структуры из содержимого текстовых документов. Приведены результаты экспериментов с применением разработанных методов на материалах русскоязычных и англоязычных корпусов, как созданных в ходе выполнения работы, так и существующих, в частности, FinTOC-2022. Представлено сравнение результатов разработанных методов с существующими системами.

В третьей главе описывается разработанный в диссертации расширяемый программный комплекс, позволяющий автоматически извлекать содержимое и восстанавливать иерархическую структуру текстовых электронных документов различных форматов.

В заключении представлены основные результаты работы.

Автореферат полностью соответствует содержанию диссертации и адекватно ее характеризует.

Диссертация и автореферат соответствуют специальности 2.3.5 Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Замечания по диссертации.

1. Желательно было провести сравнение разработанных в диссертации методов с активно развивающимися в последнее время мультимодальными языковыми моделями.
2. На странице 58 указано, что в качестве одной из моделей классификации использовался метод максимального правдоподобия (MLE), но это общий подход для оценки параметров моделей, который, в том числе, используется в логистической регрессии. Необходимо пояснить, какая именно реализация MLE применялась.
3. Для классификации строк в методе автоматического восстановления иерархической структуры выбрана библиотека XGBoost (стр. 91), при этом отсутствует сравнение как с другими моделями машинного обучения (случайный лес, нейронные сети), так и с другими реализациями градиентного бустинга (CatBoost, LightGBM).
4. В формуле для F1 на странице 58 в числителе пропущен множитель 2.
5. В работе присутствуют грамматические и пунктуационные ошибки.

Указанные замечания не влияют на общую положительную оценку диссертации.

Заключение. В целом, считаю, что диссертационная работа Беляевой Оксаны Владимировны соответствует требованиям пунктов 9–14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24 сентября 2013 года № 842, предъявляемым к кандидатским диссертациям, а её автор заслуживает присуждения учёной

степени кандидата технических наук по специальности 2.3.5 Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Официальный оппонент
доктор технических наук, доцент,
профессор Автономной
некоммерческой образовательной
организации высшего образования
«Европейский университет
в Санкт-Петербурге» Котельников Евгений Вячеславович

Подпись Котельникова Е.В. заверяю
Специалист отдела 16.03.2026г.