МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)

Кронверкский пр-т, д. 49, лит. А,
Санкт-Петербург, Россия, 197101
Тел.: (812) 480-00-00 | Факс: (812) 232-23-07
od@itmo.ru | itmo.ru

УТВЕРЖДАЮ
Проректор по научной работе
ФГАОУ ВО «Национальный исследовательский университет ИТМО»
доктор техн. наук, профессор

В.О. Никифоров
«25» марта 2025 г.

25.03.2025 № 10.09/8

ОТЗЫВ

ведущей организации 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».

1. Актуальность темы диссертационной работы.

Диссертационная работа Беляевой О. В. посвящена исследованию и разработке методов автоматического извлечения содержимого и восстановления иерархической структуры текстовых электронных документов различных форматов и предметных областей. В условиях стремительного роста объемов неструктурированных данных, исследования в области автоматической обработки и анализа многостраничных документов становятся особенно актуальными. Иерархическое представление документов позволяет системам интеллектуальной обработки оптимизировано хранить и эффективно работать с такими данными, обеспечивая их структуризацию и удобство для дальнейшего анализа.

Особую значимость приобретает развитие методов восстановления иерархической структуры документов, что особенно важно в контексте увеличения числа систем анализа и роста объемов создаваемых неструктурированных документов. Одним из наиболее распространенных форматов является PDF, автоматическая обработка которого требует

дополнительного анализа для обеспечения корректности, повышения качества и эффективности работы с такими документами.

В связи с этим, исследования, представленные в диссертационной работе Беляевой О. В., приобретают особую актуальность. В работе рассмотрены ключевые проблемы и предложены инновационные методы автоматической обработки PDF-документов, а также восстановления их иерархической структуры. Эти разработки вносят значительный вклад в развитие области автоматической обработки документов, предлагая решения, которые отвечают современным вызовам и способствуют повышению эффективности работы с большими массивами документов.

2. Научная новизна полученных результатов и выводов.

К важнейшим результатам диссертационной работы, обладающих научной новизной, относятся следующие:

1. Предложен новый метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов.
2. Предложен новый метод автоматического восстановления иерархической структуры из содержимого документов, который превосходит по качеству другие решения.

3. Достоверность полученных результатов.

Обоснованность положений диссертации подтверждается результатами проведенных экспериментов и анализом эффективности разработанных методов, а также апробацией на конференциях и научных мероприятиях всероссийского и международного уровней и научными публикациями, среди которых три индексируются в базах Scopus и Web of Science. Кроме того, получено три свидетельства о государственной регистрации программ для ЭВМ.

4. Значимость полученных результатов для науки и практики.

Представленные в диссертации результаты могут быть использованы для разработки интеллектуально-аналитических систем анализа электронных документов, в которых могут решаться задачи поиска по актуальным данным, сбор статистической информации и другие задачи анализа данных, в том числе с использованием больших языковых моделей.

Практическая ценность диссертации подтверждается актами внедрения в бюджетные и коммерческие организации. Дополнительно, стоит отметить, что разработанный программный комплекс является открытым и расширяемым для обработки новых предметных областей и форматов документов за счет разработанной архитектуры и методики расширения.

5. Структура и содержание диссертации.

Диссертация состоит из введения и трех глав, заключения, списка литературных источников из 106 наименований и 3 приложений. Общий объем работы 152 страниц текста.

Во введении обоснована актуальность темы диссертации, сформулирована цель, задачи, показаны научная новизна, практическая

ценность результатов работы, приведены основные положения, выносимую на защиту.

В первой главе приведен обзор состояния области автоматической обработки электронных документов, показан основной подход обработки, описаны основные методы извлечения содержимого документов, восстановления структуры, анализа PDF-документов, приведены результаты анализа причин некорректности PDF.

Вторая глава включает описание нового метода автоматической обработки PDF-документов с проверкой корректности текстового слоя для повышения качества и снижения времени обработки PDF. Также показана общая схема автоматической обработки документов, включающая предложенный метод. В главе также представляется новый метод восстановления иерархической структуры текстовых документов, который использует расширенное признаковое пространство и учитывает оглавление в документах. В главе приведены экспериментальные обоснования предложенных методов, которые показывают лучшее качество обработки документов на разных наборах данных.

В третьей главе описывается расширяемая архитектура программного комплекса, включающего разработанные методы. Также описана разработанная методика расширения программного комплекса. Дополнительно приводится описание технической составляющей программного комплекса: внешний интерфейс, документация, особенности запуска.

В заключении подведены итоги и обобщены результаты проведенных исследований.

6. Замечания.

По диссертации имеются следующие замечания:

1. Метод определения корректности текстового слоя обучался и тестировался на русских и английских текстах, при этом в тексте не раскрыта его применимость и расширяемость на текстах других языков.
2. В Таблице 2.1.6.3 при сравнении открытых систем также было бы полезно указать параметры запуска каждой системы.
3. Помимо описания выходного представления документа в разделе 3.5, было бы полезно привести примеры представлений документа в HTML и JSON форматах.

Тем не менее указанные замечания не снижают положительную оценку работы и не уменьшают высокой значимости полученных результатов.

7. Заключение.

Диссертационная работа Беляевой Оксаны Владимировны “Автоматическое восстановление структуры текстовых документов” является законченным научным исследованием по актуальной теме. В работе представлены результаты, имеющие важное научное и практическое значение и соответствуют специальности 2.3.5 - “Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей”, а

ее автор Беляева О.В. заслуживает присуждения ученой степени кандидата технических наук.

Диссертационная работа обсуждалась на семинаре исследовательского центра в сфере искусственного интеллекта «Сильный искусственный интеллект в промышленности» федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет ИТМО». Настоящий отзыв рассмотрен и одобрен на семинаре исследовательского центра в сфере искусственного интеллекта «Сильный ИИ в промышленности», протокол №3 от «20» марта 2025 г. На заседании присутствовали 8 человек.

Отзыв составил:
Научный руководитель исследовательского
центра в сфере искусственного интеллекта
«Сильный искусственный интеллект
в промышленности»,
доктор техн. наук Бухановский А.В.

Сведения о ведущей организации:
Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО» (Университет ИТМО).
Почтовый адрес: 197101, г. Санкт-Петербург, пр-т Кронверкский, д. 49, лит. А
Телефон: (812) 480-00-00
Веб-сайт: https://itmo.ru
Адрес электронной почты: od@itmo.ru