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

В анализе документов особая роль принадлежит этапу структуризации документов, который является ключевым и первоначальным для систем интеллектуального анализа данных, использующих современные большие языковые модели. Иерархическое представление документа позволяет разделить его на части различного уровня, что обеспечивает эффективность хранения и возможность высокоточно работать с многостраничными документами. В связи с чем, на данный момент исследования в области автоматической обработки и восстановления иерархического представления документов различного формата и типов структуры требует научного развития.

Исследования, представленные в диссертационной работе Беляевой О.В., приобретают особую актуальность в контексте роста объемов неструктурированных данных. Предложенные в работе методы и инструменты вносят значительный вклад в развитие области автоматической обработки документов, предлагая решения для эффективного анализа разнообразных документов. Также, одной из ключевых особенностей диссертационной работы является исследование проблем автоматической обработки PDF-документов, включая проверку их корректности, что подчеркивает практическую значимость и актуальность проведенных исследований в условиях повсеместного распространения документов в формате PDF.

Научная новизна

В диссертационной работе был разработан метод автоматического извлечения содержимого PDF документов с использованием проверки

текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов на русском и английском языках. Также был предложен новый метод автоматического восстановления иерархической структуры из содержимого документов, который показывает более высокое качество восстановления структуры, по сравнению с другими методами.

Практическая значимость диссертации

Практическая значимость диссертационной работы заключается в разработке методов и открытого расширяемого программного средства, позволяющего автоматически обрабатывать текстовые электронные документы с целью извлечения их содержимого и иерархической структуры. В программное средство внедрены разработанные методы автоматической обработки документов. Программное средство используется различными организациями для автоматической обработки документов, такой как восстановление структуры документов формата PDF выпускных квалификационных работ и законов, извлечение содержимого документов разного формата, в том числе изображений сканированных документов. При этом, практическая значимость диссертационной работы подтверждается актами о внедрении в несколько организаций.

Стоит отметить, что программное средство показывает эффективность обработки PDF-документов по сравнению с другими решениями, за счет разработанного метода определения корректности PDF.

Повышает практическую значимость программного средства возможность расширения поддержки новых видов документов.

Обоснованность и достоверность результатов диссертации

Достоверность результатов диссертации подтверждается проведенными экспериментами. Методы и постановка экспериментов построены корректно и согласуются с общепризнанными методологиями. Достоверность результатов не вызывает сомнений.

Апробация и публикация результатов диссертации

Результаты диссертации неоднократно обсуждались на общероссийских и международных конференциях. Результаты диссертации были опубликованы в 10 печатных изданиях, 3 из которых индексируются в базах Scopus и Web of Science.

Замечания по диссертации

1) В работе нет обоснования выбора рассматриваемых типов документов (техническое задание, нормативно-правовой акт, выпускная квалификационная работа).

2) Из текста диссертации неясно исследовался ли вопрос о том, как шум в виде различных искажений информации и ошибок в документах влияет на работу разработанных методов, в частности, на метод восстановления иерархической структуры документов.

3) В описании метода восстановления иерархической структуры документа указывается, что каждый тип (класс) строки определяется своим приоритетом в рамках определенной предметной области. Однако четко не указывается кто, как и когда определяет данные приоритеты.

4) Не рассмотрен вопрос влияния специфики русского языка при восстановлении структуры документов.

5) В работе представлено сравнение разработанного метода восстановления иерархической структуры документов с аналогами в рамках соревнования FINTOC-2022 (Таблица 2.3.3) на тестовом наборе данных, содержащего финансовые документы, где показал наилучший результат. Однако для выбранных в работе типов документов (техническое задание, нормативно-правовой акт, выпускная квалификационная работа) такое сравнение не проводилось.

Указанные замечания не влияют общую положительную оценку диссертации.

Заключение

Таким образом, диссертационная работа Беляевой О.В. является самостоятельным и завершенным научным исследованием и соответствует требованиям "Положение о порядке присуждения ученых степеней", утвержденного постановлением Правительства РФ от 24 сентября 2013 года №842, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.5 "Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей".

Официальный оппонент

Дородных Никита Олегович, 25.03.2025 г.