МЕТОДЫ ОБРАБОТКИ PDF-ФАЙЛОВ В ФИНАНСОВЫХ СИСТЕМАХ

Матлаш Т.С.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Давыдова Н.С. – канд. тех. наук, доцент

Аннотация. В статье рассмотрены современные методы обработки PDF-документов в финансовых системах, используемых для автоматизации работы с бухгалтерской отчетностью, платежными поручениями и другими финансовыми документами. Проанализированы подходы к извлечению текстовой и табличной информации, включая технологии оптического распознавания символов (OCR) для сканированных документов. Показано, что комбинированный подход, включающий структурный и семантический анализ документов, позволяет повысить эффективность автоматизации обработки финансовых данных и снизить вероятность ошибок при формировании отчетности.

Введение

Современные финансовые системы обрабатывают большие объемы документов в формате PDF, включая бухгалтерскую отчетность, счета и договоры [1]. PDF обеспечивает сохранение визуальной структуры документа, но отсутствие единой логической разметки затрудняет автоматизированное извлечение данных. Повышение точности и скорости обработки финансовых документов требует применения современных методов анализа и структурирования информации [2].

Основная часть

Существующие методы обработки PDF включают:

– извлечение текстового слоя и структурный анализ документа;

– распознавание таблиц и объединённых ячеек;

– применение OCR для сканированных документов;

– использование регулярных выражений и семантического анализа для выделения ключевых реквизитов: даты, суммы, валютные показатели, сведения о контрагентах [3].

Анализ программных средств показал, что большинство решений ориентировано на выполнение отдельных задач и не обеспечивает универсальности для работы с различными типами финансовых документов. Это создает необходимость разработки адаптивного метода, позволяющего эффективно извлекать ключевые реквизиты и структурировать данные независимо от формата документа [4].

В рамках исследования предложен метод обработки PDF-документов, включающий:

– классификацию типа документа;

– предварительную обработку текста и таблиц;

– семантический анализ данных;

– использование регулярных выражений для выделения ключевых реквизитов, таких как даты, суммы, валютные показатели и сведения о контрагентах.

Для оценки эффективности метода применялись стандартные метрики информационного поиска – точность и полнота извлечения данных. Результаты показали высокую точность обработки, устойчивость метода к разнообразию форматов документов и возможность автоматизации обработки больших объемов финансовой информации [5].

Использование предложенного метода позволяет:

– ускорить обработку финансовых документов;

– снизить вероятность ошибок при формировании отчетности;

– повысить качество аналитических данных;

– интегрировать обработку документов в финансовые информационные системы.

В перспективе дальнейшее развитие метода может включать машинное обучение для автоматической классификации и семантического анализа более сложных структур документов, что расширит область применения и повысит адаптивность метода.

Заключение

Проведен анализ методов обработки PDF-документов в финансовых системах. Выявлены преимущества и ограничения современных технологий, включая OCR, структурный и семантический анализ. Предложенный комбинированный подход позволяет автоматизировать обработку финансовых документов, повышает точность извлечения данных и снижает риск ошибок. Дальнейшее развитие может включать интеграцию методов машинного обучения для работы с более сложными структурами PDF и расширения области применения.

Список использованных источников:

1. Loughran T., McDonald B. Textual Analysis in Accounting and Finance.

2. Li F. The Information Content of Forward-Looking Statements.

3. Smith R. An Overview of the Tesseract OCR Engine.

4. Shinyama Y. PDF Document Analysis for Information Extraction.

5. Klampfl S. Unsupervised Document Structure Analysis.