

ОТЗЫВ

на автореферат диссертационной работы Шигарова Алексея Олеговича «Методы и инструментальные средства автоматизации процессов извлечения данных из таблиц электронных документов неструктурированного формата», представленной на соискание ученой степени доктора технических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Рассматриваемая диссертационная работа посвящена важной и востребованной научной проблеме – разработке методов и инструментальных средств для комплексной автоматизации процессов распознавания, анализа и интерпретации документных таблиц произвольной формы, являющихся неструктурированными данными, заключенными в электронных документах.

Актуальность избранной темы диссертации не вызывает сомнений. В условиях повсеместного распространения технологий анализа больших данных проблема извлечения структурированной информации из машинно-ориентированных, но неструктурированных и слабоструктурированных форматов, таких как PDF и Excel, является одной из ключевых. Автор справедливо отмечает, что ручной труд в этой области неэффективен, а существующее многообразие форм представления таблиц требует создания гибких и универсальных решений. Таким образом, направление исследований, выбранное соискателем, является своевременным и значимым для развития области программной инженерии и анализа данных.

Научная новизна работы, как следует из автореферата, соответствует положениям (пп. 1, 2, 3) паспорта специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Предложенные автором проблемно-ориентированные модели и методы для обработки таблиц произвольной структуры вносят существенный вклад в область математического и программного обеспечения вычислительных систем.

Практическая значимость результатов очевидна и подтверждается разработкой автором конкретных программных продуктов – TabbyPDF и TabbyXL. Создание таких инструментальных средств позволяет существенно упростить и ускорить разработку прикладного программного обеспечения для извлечения данных, что повышает результативность работы аналитиков в различных предметных областях.

Достоверность полученных результатов обоснована приведенными экспериментальными данными, а также качественным и количественным сравнением с существующими аналогами, что свидетельствует о высокой эффективности предложенных решений.

Положительно следует отметить основательную проработку материала – от исторического обзора и анализа существующих проблем до детализации собственной технологии, включающей как методическую, так и инструментальную составляющие. Апробация работы и объем публикаций являются более чем достаточными и соответствуют требованиям ВАК.

Работа оценивается положительно, но при чтении автореферата можно сделать несколько вопросов и замечаний:

Хотя работа демонстрирует эффективность и преодолевает многие ограничения существующих аналогов, в самом автореферате имеются определенные ограничения, касающиеся производительности и полноты учета приемов оформления:

1. Ограничения по качеству на сложных коллекциях: На предметно-ориентированных тестах Troy200 и SAUS200 (включающих таблицы государственной статистики со сложной компоновкой) качество работы тестируемых решений не достигает 100 %. Это объясняется противоречивостью приемов оформления таблиц.

2. Неполный учет приемов оформления иерархических заголовков: При оценке производительности на коллекции SAUS200, где использовались разнообразные приемы оформления иерархических заголовков (отступы,

шрифты, выравнивание), большая часть допущенных ошибок была вызвана тем, что не были учтены все варианты применения перечисленных приемов.

3. Незначительное отставание от промышленных аналогов в РТ: На предметно-независимом тесте ICDAR-2013 решение TabbyPDF, хотя и показывает результаты, близкие к лучшим академическим аналогам, немного уступает некоторым индустриальным продуктам.

4. Фокус на НПОД/ПК: В качестве источников поддерживаются НПОД (PDF) и ПК (Excel/Sheets), хотя другие форматы с машиночитаемым текстовым содержимым могут быть преобразованы к поддерживаемым.

Однако, по-нашему мнению основное качественное преимущество работы заключается в преодолении главного ограничения современных методов — отсутствия поддержки свободной компоновки таблиц и структурированности ячеек.

Отмеченные замечания не снижают научную и практическую ценность диссертационной работы.

В целом рассматриваемая диссертация представляет собой современное, актуальное и практически значимое исследование. На основании изучения автореферата, считаю, что диссертационная работа Шигарова Алексея Олеговича соответствует требованиям ВАК РФ для докторских диссертаций, а ее автор заслуживает присуждения искомой ученой степени доктора технических наук по специальности 2.3.5. – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Отзыв представил: Потапов Вадим Петрович, доктор технических наук, профессор, главный научный сотрудник, заведующий лабораторией геоинформационного моделирования Кемеровского филиала Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ).

Я, Потапов Вадим Петрович, даю свое согласие на включение своих персональных данных в документы, связанные с работой диссертационного совета, их дальнейшую обработку и передачу в соответствии с требованиями Минобрнауки России.

Подпись: _____ 

В.П. Потапов

Дата: «25» ноября 2025 г.

Контактная информация:

ФИЦ ИВТ:

630090, г. Новосибирск, пр. Академика Лаврентьева, 6; (3833) 30-61-50;
ict@ict.nsc.ru; <http://www.ict.nsc.ru>

Кемеровский филиал ФИЦ ИВТ:

650025, г. Кемерово, ул. Рукавишникова, 21;
(3842) 21-14-00; kem@ict.nsc.ru;
<http://www.ict.nsc.ru/ru/structure/branches/kemerovo-branch>

Подпись доктора технических наук, профессора Потапова Вадима Петровича, заверяю:

Ученый секретарь Федерального исследовательского
центра информационных и вычислительных
технологий, к.ф.-м.н.



Н. В. Киланова

