

## ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

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

**Актуальность диссертационного исследования** связана с наличием в мире большого массива табличных данных в различных документах (научных статьях, отчетах, формах и др.), которые представляют собой ценный источник информации в различных предметных областях. Данные, извлеченные из документных таблиц, используются в системах поддержки принятия решения, системах, основанных на знаниях, формировании онтологий предметных областей и др. Однако, такие таблицы, как правило, предназначены для понимания человеком и не имеют явной семантики (формальной модели данных), что затрудняет их автоматическую интерпретацию компьютерными программами. Документные таблицы представляют информацию как комбинацию текста и пространственного расположения его фрагментов, что требует специальных методов извлечения данных, основанных на методах распознавания структуры и состава первоисточников.

Проблема автоматизации извлечения данных из таблиц стала интенсивно исследоваться в последние два десятилетия. Поэтому до сих пор отсутствует единое понимание границ данной проблематики, именуемой в литературе как «автоматизированное понимание таблиц» (АПТ), унифицированная терминология и комплексные решения, охватывающие весь процесс АПТ — от обнаружения таблицы в документе до ее семантической интерпретации с помощью внешних баз знаний. Таким образом, исследования в этой области критически важны для раскрытия потенциала больших массивов табличных данных для развития систем искусственного интеллекта.

**Научная новизна работы** определена следующими основными результатами:

- Разработана единая схема задач АПТ, которая устраняет терминологические противоречия между смежными областями (компьютерное зрение, управление данными, информационный поиск). Благодаря согласованной терминологии и многоуровневой детализации, данная схема точнее отражает современную парадигму АПТ по сравнению с существующими подходами.
- Предложен метод распознавания таблиц, который использует особенности представления страниц нередактируемых печатно-ориентированных документов (НПОД) в формате языков описания страниц PDL (Page Description Language). В отличие от аналогов, метод целенаправленно задействует PDL-специфичную информацию для анализа компоновки страницы, что позволило реализовать новые методики автоматического обнаружения и сегментации таблиц, которые также можно использовать совместно с предварительно обученными искусственными нейронными сетями для улучшения качества результатов распознавания таблиц.
- Создана гибкая модель документных таблиц, которая не накладывает жестких ограничений на их компоновку, структуру ячеек или иерархию заголовков. В отличие от моделей, используемых в существующих решениях, данная модель способна корректно представлять таблицы произвольной сложности, что расширяет область применения методов АПТ.
- Разработан метод, позволяющий анализировать и интерпретировать таблицы с произвольной структурой: свободной компоновкой, ячейками с наборами разнотипных значений, многоуровневыми заголовками. Его ключевое отличие от аналогов — основа на подходе, позволяющем пользователям программировать логику обработки с помощью правил, что обеспечивает гибкость и адаптивность.

- Создан специализированный язык (DSL) правил, который позволяет пользователям описывать логику извлечения данных из таблиц. В отличие от языков общего назначения, этот DSL позволяет разработчику сосредоточиться исключительно на задачах анализа и интерпретации, тем самым упрощая создание соответствующего программного обеспечения.

- На основе предложенных решений разработан комплекс инструментов для автоматического извлечения данных из таблиц в таких форматах как PDF, Excel или других, приводимых к ним. Его конкурентные преимущества состоят в следующем: при распознавании таблиц в PDF используется анализ компоновки страницы и фильтрация кандидатов для повышения точности. При извлечении данных реализован принцип разделения ответственности: правила анализа и интерпретации логически отделены от моделей данных и алгоритмов обработки, что повышает гибкость и удобство сопровождения системы.

**Теоретическая и практическая ценность работы.** Теоретические и технологические решения, разработанные в рамках диссертации, могут быть положены в основу создания новых информационных технологий и программных продуктов сбора и интеграции данных из документных таблиц. Сегодня, сталкиваясь с задачами АПТ на практике, предметные специалисты и разработчики программного обеспечения (ПО) обычно прибегают к инструментальным средствам общего назначения, предлагая собственные реализации однотипных задач. В сравнении с последними, предлагаемые решения могут позволить сократить затраты на разработку целевого ПО. В сравнении с имеющимися инструментальными средствами общего назначения, предлагаемые в диссертации решения могут позволить сократить затраты на разработку целевого ПО АПТ. Последнее особенно важно в тех случаях, когда необходимо в короткие сроки и при недостатке ресурсов разработать ПО для массовой обработки табличных данных. Вышеизложенное согласуется со Стратегией научно-технологического развития Российской Федерации по приоритету: «создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта».

**Обоснованность и достоверность полученных результатов** подтверждается:

- корректным применением выбранных методов исследования;
- значительным экспериментальным материалом и сравнением с существующими аналогами;
- сравнительным анализом с существующими аналогами, продемонстрировавшим преимущества предложенных решений;
- успешным опытом внедрения разработанных методов и инструментов в рамках разнопредметных научных и промышленных проектов.

**Публикации и апробация.** Основные результаты диссертации представлены в 70 работах: 12 научных статей в журналах, рекомендованных ВАК (все они относятся к первой категории K1), 34 статей в международных рецензируемых изданиях, индексируемых WOS/Scopus, шесть свидетельств о государственной регистрации программ для ЭВМ. Результаты исследований были неоднократно представлены и получили одобрение на международных и всероссийских конференциях, а также на научных семинарах ведущих российских и зарубежных организаций. Таким образом, количество и высокий уровень публикаций полностью соответствуют требованиям к докторской диссертации по техническим наукам.

**Личный вклад** автора в диссертационную работу подтверждается материалами диссертации и автореферата и включает: постановку научных задач; разработку методов, моделей, языковых и инструментальных средств; проведение вычислительных экспериментов и анализ полученных данных. Вклад автора в совместные публикации является конкретным и четко обозначенным. Все используемые материалы снабжены необходимыми цитатами и ссылками на источники.

**Замечания по диссертации.** По представленной диссертационной работе имеются следующие замечания:

Замечание 1. На странице 104 присутствует следующее утверждение: "В этом случае она является строкой некоторого табличного региона *тогда и только тогда*, когда сумма ширин всех ее пробельных промежутков строки относительно ширины самой строки не пре-

вышает заданной пороговой величины:..." Такие утверждения должны сопровождаться доказательством необходимости и достаточности условий, что в диссертации не сделано.

Замечание 2. В качестве одного из целевых представлений в диссертации определена база данных. Однако автор, при этом, ограничивается условиями первой нормальной формы. Это не достаточно для определения не избыточного и непротиворечивого множества классов объектов, представленных в итоговых таблицах. Исследование зависимостей в данных и последующая нормализация таблиц до пятой нормальной формы позволила бы решить указанную проблему.

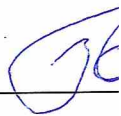
Замечание 3. Многомерная таблица в работе определяется как отдельная сущность. Однако, исследования, проводимые под руководством автора отзыва, показывают, что многомерная таблица является трансформацией нормализованной базы данных.

Замечание 4. При проведении тестирования в работе используется набор стандартных тестов. Целесообразно показать использование тестов, на которых происходила отладка программ, и которые показывают существенное преимущество разработанной технологии.

Приведенные в настоящем отзыве замечания не снижают научную и практическую ценность представленных соискателем результатов диссертационного исследования и не влияют на общую положительную оценку данной работы.

**Заключение.** Диссертация является самостоятельным, целостным исследованием, содержащим новые научные результаты. Разработанный автором комплекс научно обоснованных решений вносит значительный вклад в области информационных технологий, повышая эффективность и надежность процессов обработки данных и знаний, представленных в виде документных таблиц, в вычислительных системах. Работа полностью соответствует установленным требованиям к диссертациям на соискание ученой степени доктора наук «Положения о порядке присуждения ученых степеней», а ее автор, Шигаров Алексей Олегович, заслуживает присуждения ученой степени доктора технических наук по специальности 2.3.5. «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент: \_\_\_\_\_



— Сергей Владимирович Зыкин

«10» ноября 2025 г.

Доктор технических наук, профессор.

Заведующий лабораторией методов преобразования и представления информации Омского филиала Федерального государственного бюджетного учреждения науки Института математики им. С.Л. Соболева Сибирского отделения Российской академии наук,

Адрес организации: 644099, г. Омск, ул. Певцова, 13.

Телефон: +7 (3812) 23-67-39

Электронная почта: [zykin@ofim.oscsbras.ru](mailto:zykin@ofim.oscsbras.ru)

Подпись С.В. Зыкина удостоверяю:

ИО ученого секретаря ОФ ИМ



ф



Раиса Хиссатовна Багаутдинова