ОТЗЫВ

на автореферат диссертационной работы 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»

Диссертационная работа А. О. Шигарова посвящена актуальной научной проблеме – разработке методов и инструментальных средств для автоматизации извлечения данных из произвольных документных таблиц, представленных в неструктурированном виде. Актуальность темы обусловлена широким распространением такой информации, наличием необходимости структурировать ее в различных прикладных задачах, и отсутствием универсальных решений, обеспечивающих такое структурирование.

Научная новизна и теоретическая значимость работы заключаются в следующих основных результатах, выносимых на защиту:

- Предложена усовершенствованная структура задач для области «автоматизированного понимания таблиц». В ней систематизирована терминология из смежных исследовательских направлений. Благодаря согласованной терминологии и многоуровневой декомпозиции, новая система задач точнее отражает суть процессов по сравнению с существующими подходами.
- Разработан новый метод автоматизации процессов распознавания таблиц в нередактируемых печатно-ориентированных документах. Его главное новшество состоит в использование специфичной для языка описания страниц информации. Показано, как эти данные применяются для анализа макета страницы, обнаружения таблиц и их сегментации.
- Создана гибкая модель таблицы для представления табличной информации в процессах «автоматизированного понимания». В отличие от известных аналогов, она не накладывает жестких ограничений на структуру, поддерживая таблицы с произвольной компоновкой, вложенными ячейками и многоуровневыми заголовками.
- Предложен метод автоматизации процессов анализа и интерпретации редактируемых таблиц. Впервые этот процесс реализован с помощью настраиваемых пользовательских правил. Метод обеспечивает полную поддержку сложных структур таблиц, включая свободное расположение элементов, составные ячейки и иерархические заголовки.
- Разработан проблемно-ориентированный язык, предназначенный для написания правил анализа и интерпретации таблиц. В отличие от языков общего назначения, он позволяет разработчику сосредоточиться исключительно на логике извлечения данных, а не на технических деталях реализации.
- На основе теоретических наработок создан комплекс инструментальных средств. Его ключевыми функциями являются конвертация: (1) преобразование таблиц из печатных документов (PDF) в редактируемый формат с возможностью последующей ручной коррекции; (2) автоматическое извлечение наборов записей в канонической форме из редактируемых таблиц с помощью пользовательских правил. По сравнению с существующими аналогами, данный комплекс в целом предоставляет «сквозную» функциональность для работы с табличными данными.

Практическая значимость подтверждена внедрением разработанного программного обеспечения в пять научных и три индустриальных проекта для анализа технических, финансовых документов, интеграции данных и создания систем, основанных на знаниях. Результаты работы получены в рамках выполнения государственных заданий и грантов (РНФ, РФФИ и др.). Автором в составе коллектива получено 6 свидетельств о регистрации программ для ЭВМ.

Достоверность результатов подтверждается корректным использованием методов, работоспособностью программного обеспечения, экспериментальными данными и сравнением с аналогами. Основные положения диссертации опубликованы в 34 статьях в изданиях WOS/Scopus и 12 статьях в научных журналах, рекомендуемых ВАК РФ.

В качестве недостатка автореферата можно отметить, что представленные в таблицах 1–3 и 5–6 результаты численных экспериментов не демонстрируют в явном виде преимущества разработанных методов по сравнению с аналогами. При беглом ознакомлении это может привести читателя к некорректной качественной оценке эффективности предложенного подхода. И, хотя в разделе «Выводы к пятой главе» перечислены ключевые достоинства разработанного метода, их стоило бы подчеркнуть более явно и связать с экспериментальными данными.

Указанное замечание не снижает значимость результатов диссертационной работы и не влияет на ее положительную оценку.

Диссертация А. О. Шигарова является завершенной научной работой, соответствует требованиям ВАК РФ и рекомендуется к защите на соискание ученой степени доктора технических наук по специальности 2.3.5. – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Отзыв представили:

Аветисян Арутюн Ишханович, Директор Института системного программирования РАН, Доктор физико-математических наук, академик РАН.
Согласен на включение своих персональных данных, включённых в настоящий отзыв, в документы, связанные с работой диссертационного совета, а также их дальнейшую обработку и передачу в соответствии с требованиями Минобрнауки России.

А. И. Аветисян
«26» ноября 2025 г.

Турдаков Денис Юрьевич, Заведующий отделом информационных систем Института системного программирования РАН, Кандидат физико-математических наук.
Согласен на включение своих персональных данных, включённых в настоящий отзыв, в документы, связанные с работой диссертационного совета, а также их дальнейшую обработку и передачу в соответствии с требованиями Минобрнауки России.

Д. Ю. Турдаков
«26» ноября 2025 г.

Контакты:
Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В. П. Иванникова Российской академии наук (ИСП РАН).
109004, г. Москва, ул. А. Солженицына, дом 25.
Тел: +7(495) 912-44-25, эл. почта: info-isp@ispras.ru.