

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 темы диссертации

Опубликованные в Интернете и в печатном виде документы содержат огромные объемы полезной информации. Качественное извлечение этой информации, представляет сложную техническую и научную задачу. Предметом рассматриваемой диссертации является извлечение важной части информации - таблиц, содержащихся электронных документах.

Таблицы форматов PDF, HTML, Excel и др., известные как «документные таблицы», представляют собой глобальный масштабный источник реляционных данных. Общемировой объем этой информации огромен и востребован. Эти данные являются критически важным ресурсом для систем принятия решений, интеллектуального анализа данных, построения баз знаний и информационного поиска.

Для использования этих данных их необходимо преобразовать в машиночитаемую структурированную форму (базы данных, графы знаний). Основная сложность заключается в огромном разнообразии компоновок, форматов и стилей таблиц, а также в отсутствии у них формальной схемы данных, что делает невозможным прямое применение стандартных методов обработки текста и структурированных данных. Эта проблема, остается нерешенной несмотря на более чем 30-летнюю историю исследований.

Представленный в диссертации анализ современной литературы и коммерческих продуктов показывает, что все существующие решения являются частными и опираются на предположения о «типичной» структуре таблиц. Они работают лишь с ограниченным набором шаблонов (простые заголовки, атомарные ячейки), игнорируя огромный пласт таблиц с произвольной и сложной структурой. Это существенно сужает область их применимости.

Таким образом, разработка методов и алгоритмов для автоматического извлечения данных из документных таблиц произвольной структуры является актуальной научной проблемой. Ее решение имеет высокую практическую значимость для обработки ключевых типов документов: нередатируемых

печатно-ориентированных документов (PDF, PostScript и др.) и рабочих книг (Excel, Google Sheets и др.).

Обоснованности и достоверность научных результатов диссертации

Обоснованность научных результатов, выводов и рекомендаций, сформулированных в диссертации, базируется на всестороннем анализе современных тенденций в области автоматизированного понимания табличной информации. Данное направление раскрыто автором как ключевое для повышения эффективности разработки цифровых продуктов на основе извлечения реляционных данных из документных таблиц. Актуальность, цель, задачи, научная новизна, теоретическая и практическая значимость работы детально аргументированы.

Во многом работа базируется на глубоком и содержательном обзоре современного состояния исследований, в котором не только проанализирована предметная область, но и выявлены достижения, проблемы и интересные перспективы. Общая задача автоматизированного понимания таблиц решалась через этапы: распознавание таблиц, анализ таблиц, интерпретация таблиц. В каждом из этих направлений были сделаны модельные построения, сформированы технические и алгоритмические решения. В своем исследовании, автор опирался не только на свои модели, но и на современные результаты в областях реляционных таблиц, структур документов, машинном обучении, задач классификации, ролевого анализа,

Имеются и заделы на будущее. В частности, в обзорной части рассмотрена задача семантического аннотирования, что является хорошей перспективой к выходу на построения Semantic Web и интеграцию баз знаний.

Итоговый программный комплекс базируется на выполненных и описанных в работе теоретических построениях с использованием качественных программ свободного распространения, решающих локальные подзадачи. На этом и на проведенных испытаниях и экспериментах базируется обоснованность и полезность программ, главные из которых зарегистрированы и прошли проверку во внешне проведенных больших работах. В программном комплексе широко использованы современные стандарты, придающие ему гибкость и позволяющие осуществлять последующее развитие.

Достоверность научных результатов диссертационного исследования обеспечивается согласованностью полученных данных с ранее известными научными положениями, корректным применением современных методов исследования, адекватных поставленным задачам; репрезентативностью экспериментальных и сравнительных данных, апробацией ключевых

положений работы на международных и всероссийских конференциях, публикацией основных результатов в рецензируемых отечественных и зарубежных изданиях, включая журналы с высоким импакт-фактором, анализом результатов обработки в сравнении с известными открытыми и коммерческими решениями, внедрением полученных результатов в практическую деятельность.

Новизна научных результатов диссертации

Новизна работы заключается в комплексном решении научной проблемы автоматизированного извлечения реляционных данных из таблиц с произвольной структурой, представленных с помощью языков описания страниц и форматов рабочих книг. К новым научным результатам, полученным автором диссертации, следует отнести следующие:

1. Предложена новая структура совокупности задач «автоматизированного понимания табличной информации», которая может применяться для проектирования программных систем извлечения данных из произвольных документных таблиц.
2. Новый подход к распознаванию таблиц для нередактируемых печатно-ориентированных документов с использованием специфики представления таблиц языками описания страниц.
3. Впервые предложен метод автоматизации анализа и интерпретации таблиц с произвольной структурой, представленных в редактируемом формате рабочих книг, на основе трансляции и исполнения пользовательских правил.
4. Предложена новая модель представления документных таблиц, применимая к широкому кругу приложений за счет поддержки произвольности табличной структуры.
5. Впервые создан проблемно-ориентированный язык программирования для разработки программ анализа и интерпретации таблиц в виде наборов правил.
6. Разработан новый комплекс инструментальных средств, предназначенный для упрощения разработки цифровых продуктов (сервисов, баз данных, графов знаний и др.) на основе автоматизации извлечения данных из произвольных таблиц, представленных в электронных документах неструктурированного формата.

Замечания по диссертации

1 В обзоре предметной области явным образом указано, что “Целевым представлением извлекаемой информации обычно является либо база данных, либо граф знаний [стр. 24]”, есть указания о том, что можно связать

извлечение таблиц со словарями (спецификациями), семантическими базами данных, связанными данными. Однако, эти темы в исследовании не развиваются, что вызывает вопросы, хотя и вполне объяснимые. Аналогичные моменты проявляются в подразделе “Семантическое аннотирование” [стр. 51].

2 Не очень отчетливо описана работа с таблицами, разорванными (в диссертации используется термин “перерезы”) на несколько частей как по “горизонтали” (по набору полей), так и по “вертикали”, т.е. по набору строк.

3 Для анализа и интерпретации таблиц, в работе предложен язык CRL задания правил. При том, что модель документной таблицы, а также синтаксис CRL подробно описаны, семантику правил понять трудно. Возможно помогло бы включение в раздел 3 объяснение функционирования правил на простых примерах. Правда, в 4-м разделе произведен возврат к правилам, даны примеры разработки CRL-правил, что частично компенсирует указанные обстоятельства.

4 Программы были опробованы и протестированы на указанных наборах предметно-ориентированных тестов. Стоило бы охарактеризовать эти тесты: какие особенности имеются в извлекаемых таблицах, какие трудности надо преодолевать для извлечения, какие объемы и сложность решаемых задач.

Заключение по диссертации

Представленная диссертация является научно-квалификационной работой, в которой решена научная проблема, имеющая важное хозяйственное значение, а именно упрощение разработки цифровых продуктов (сервисов, баз данных, графов знаний и т. д.) на основе автоматизации извлечения реляционных данных из документных таблиц с поддержкой произвольности табличной структуры. Важное хозяйственное значение данной научной проблемы обосновывается существенным и постоянным ростом объемов табличной информации, представленной в неструктурированном виде, и необходимостью приведения ее к структурированному виду при реализации приложений поддержки принятия решений, аналитики данных, информационного обмена и других, необходимостью интеграции данных из разных источников для комплексного решения научно-технических и народнохозяйственных задач.

Диссертация написана автором самостоятельно, вопросы авторства и соавторства аккуратно рассмотрены в разделе “Личный вклад автора”. Будучи посвященной одной цели, а именно созданию методов и инструментальных средств автоматизации процессов распознавания, анализа и интерпретации документных таблиц, позволяющих упростить разработку цифровых продуктов на основе извлечения реляционных данных из документных

таблиц, представленных в неструктурированном виде, диссертация обладает внутренним единством. Автором диссертации выдвигаются для защиты новые научные результаты, направленные на повышения эффективности автоматизации процессов извлечения данных из таблиц электронных документов неструктурированного формата. Выдвигаемые для защиты результаты соответствуют направлениям исследований паспорта специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Диссертация свидетельствует о личном вкладе ее автора в науку: ему принадлежат основные научные результаты, выдвигаемые для защиты, за исключением результата по разработке комплекса инструментальных средств для автоматизации процессов извлечения данных из произвольных таблиц, представленных в документах неструктурированного формата, в котором программная реализация выполнена в неделимом соавторстве.

В диссертации приводятся сведения о практическом использовании полученных автором диссертации научных результатов в научной и индустриальной сфере. Разработанные методы и инструментальные средства нашли применение в восьми проектах (из них пять проектов выполнено сторонними коллективами) при решении прикладных задач анализа документов (технических, финансовых и научных), интеграции данных (государственной статистики и медиапланирования), конструирования онтологии предметной области и кросс-контекстного обмена бизнес-документами. Соответствующие копии подтверждений внедрения результатов в индустриальной сфере представлены в приложении к диссертации.

В диссертации приводятся результаты оценки производительности предложенных автором диссертации решений автоматизации процессов распознавания, анализа и интерпретации таблиц на нескольких известных тестах (ICDAR, SciTSR, IAIS, Troy, SAUS). Автором представлены количественные и качественные сравнения решений, предложенных им в диссертации, с другими известными решениями (Tango7, Senbazuru8, HUSS).

Основные научные результаты диссертации опубликованы в 12 рецензируемых научных изданиях, рекомендованных ВАК (все 12 изданий относятся к первой категории итогового распределения журналов Перечня ВАК по категориям в 2023 году, а четыре из них — к первому уровню «Белого списка»). Для шести программ для ЭВМ, разработанных в рамках диссертационного исследования, получены свидетельства о государственной регистрации в Роспатенте.

В диссертации соискатель ученой степени корректно ссылается на источники использованных им материалов. В диссертации приводится описание личного вклада соискателя ученой степени и вклада его соавторов в результаты совместных научных работ.

Диссертационная работа соответствует требованиям «Положения о присуждении ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.2013 № 842 (ред. от 25.01.2024), предъявляемым к диссертациям на соискание ученой степени доктора технических наук, 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:

Марчук Александр Гурьевич,

Доктор физико-математических наук, профессор,

Заведующий лабораторией информационных систем Института систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН).

Контактная информация:

Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова СО РАН, 630090, г. Новосибирск, проспект Лаврентьева, 6.

Телефон: +7 (383) 330-86-52, эл. почта: iis@iis.nsk.su

Марчук Александр Гурьевич

«17» ноября 2025 г.

Личную подпись завер
Начальник
отдела кадров

20.11.2025



Шигаров О.А.