

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»

Актуальность темы исследования. Диссертационная работа Шигарова А.О. посвящена решению важной проблемы – разработке методов автоматизации процессов извлечения структурированных данных из документных таблиц неструктурированного формата. Актуальность темы исследования обусловлена ростом объемов табличной информации в электронных документах при недостаточной эффективности существующих методов обработки произвольных табличных структур.

Структура и содержание работы. Диссертация состоит из введения, шести глав, заключения, списка литературы и приложений. Логика изложения материала стройная, каждая глава последовательно раскрывает этапы исследования. Глава 1 содержит анализ исследуемой проблематики и современных решений. Главы 2–3 предлагают новые методы распознавания, анализа и интерпретации таблиц. Глава 4 описывает разработанные инструментальные средства. Глава 5 представляет экспериментальную оценку и сравнение с аналогами. Глава 6 демонстрирует практическое применение результатов. Объем работы (291 с. без приложений) соответствует рекомендациям ВАК по техническим наукам, а использование 436 источников литературы отражает полноту исследования темы.

Научная новизна. Автором получен ряд новых результатов – предложен комплекс методов и моделей для автоматизации процессов извлечения данных из документных таблиц с произвольной структурой. Прежде всего, унифицирован терминологический базис «автоматизированного понимания таблиц», объединивший области компьютерного зрения, управления данными и информационного поиска. Для решения задачи распознавания таблиц в нередактируемых печатно-ориентированных документах (представленных на языке описания страниц – PDL) разработан метод, основанный на правилах анализа компоновки страницы. Этот метод использует низкоуровневые свойства PDL-представления документа, что позволило создать эффективные методики сегментации страниц, обнаружения таблиц и распознавания их структуры. Для представления сложных таблиц разработана гибкая модель, которая, кроме типичных компоновочных свойств, учитывает ряд таких особенностей, как свободная компоновка, ячейки с собственной внутренней структурой и многоуровневые иерархические заголовки. На базе этой модели диссертантом создан метод анализа и интерпретации таблиц, основанный на исполнении пользовательских правил. Для удобства конечных пользователей разработан новый проблемно-ориентированный язык, позволяющий описывать правила извлечения, анализа и интерпретации таблиц под конкретные запросы. Теоретические положения поддержаны комплексом программно-инструментальных средств, автоматизирующим основные процессы извлечения данных из документных таблиц с произвольной структурой.

Обоснованность и достоверность полученных результатов обусловлены использованием корректных методов исследования. Разработанные методы, методики, модели и инструментальные средства апробированы публикациями в рецензируемых журналах и докладами на профильных конференциях. Основные результаты получены в рамках выполнения ряда научных проектов, поддержанных РФФИ, РНФ и Советом по грантам Президента РФ, с 2012 по 2023 год, где автор участвовал в качестве руководителя. Эффективность предлагаемых методов подтверждается представленными экспериментальными оценками на существующих текстовых коллекциях, а также их количественным и качественным сравнением с конкурентными решениями.

Апробация результатов диссертации представлена 12 статьями в журналах, рекомендуемых ВАК и 34 статьями в изданиях, индексируемых международными базами цитирования WoS/Scopus. Основные результаты докладывались на международных и всероссийских конференциях (в т.ч. ACM DocEng, ICIST, DAMDID/RCDL), а также семинарах ИДСТУ СО РАН, ОНИТ СО РАН и других научных организаций.

Теоретическая и практическая значимости результатов работы. Теоретическая значимость работы состоит в формировании основ для разработки программного обеспечения, упрощающего обработку таблиц произвольной структуры. Практическая ценность работы подтверждена созданием комплекса инструментальных средств, реализующих предложенные автором методы, а также внедрением результатов в научные (в ИДСТУ СО РАН, Институте национального развития Монголии, Гуанчжоуском университете, Пекинском технологическом институте) и промышленные проекты (в компаниях «CloudSDS Inc.», «Mosaik Risk Solutions Inc.», «ООО Адвентум Консалтинг»). Имеются акты о внедрении результатов диссертационного исследования. Полученные результаты имеют значительный потенциал для применения в системах интеллектуального анализа и интеграции данных, а также управления знаниями.

Замечания по диссертационной работе:

1. В автореферате нет упоминаний о применимости активно развивающихся больших языковых моделей (LLM) к задачам обработки документных таблиц.
2. Приведенные в автореферате случаи практического применения полученных результатов охватывают либо только распознавание, либо только анализ и интерпретацию таблиц, но не оба указанных аспекта в комплексе.
3. На отдельных задачах автоматизированного понимания таблиц конкурентные решения дают лучшее качество работы по сравнению с предлагаемыми автором.
4. В автореферате имеется ряд неточностей. Например, Пекинский технологический институт не вполне верно указан как «Пекинский технологический университет» (с. 34).

Перечисленные замечания не оказывают существенного влияния на ценность и значимость основных результатов выполненного исследования.

Заключение. В диссертационной работе Шигарова А.О. получены новые научные результаты в области автоматизации извлечения данных из произвольных документных таблиц, которые имеют существенную научную новизну, теоретическую и практическую значимость. Судя по автореферату, представленная на отзыв диссертация соответствует научной специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» и действующим требованиям ВАК РФ к докторским диссертациям, а ее автор, Шигаров А.О., заслуживает присуждения ученой степени доктора технических наук по указанной специальности.

На включение персональных данных из настоящего отзыва в документы, связанные с работой диссертационного совета, а также их дальнейшую обработку, согласен.

Профессор Департамента
бизнес-информатики НИУ ВШЭ,
Доктор технических наук, доцент



Зыков Сергей Викторович

«28» октября 2025 г.

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики».
119049, г. Москва, Шаболовка ул., д. 28/11, стр. 9, каб. 1215.
Тел: +7 (495) 772-9590 * 28240, эл. почта: szykov@hse.ru

Подпись лица, предоставившего отзыв, удостоверяю

СПЕЦИАЛИСТ ПО КАДРОВОМУ ДЕЛУ

КИСЛИНСКАЯ П.А.

