

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. В современных условиях наблюдается экспоненциальный рост объемов информации, представленной в электронных документах неструктурированного формата. Значительная часть этих данных организована в виде таблиц, которые содержатся в разнородных источниках, таких как печатно-ориентированные документы (PDF, PostScript), веб-страницы (HTML) и рабочие книги (Excel, Sheets). Эти так называемые документные таблицы являются ценным ресурсом для систем поддержки принятия решений, интеллектуального анализа данных, конструирования баз знаний и информационного поиска.

Однако извлечение структурированных данных из таких таблиц сопряжено со значительными трудностями, обусловленными большим разнообразием способов их компоновки, форматирования и наполнения, а также отсутствием формальной схемы данных. Существующие методы и инструменты автоматизированного понимания таблиц (АПТ) в основном ориентированы на ограниченный набор типичных структур и не поддерживают произвольность табличной организации, такую как свободная компоновка, структурированность ячеек и иерархичность заголовков. Это существенно сужает область их применимости и ограничивает возможности массовой обработки табличной информации, особенно в контексте нередатируемых печатно-ориентированных документов (НПОД) и рабочих книг (РК).

Таким образом, разработка новых методов и инструментальных средств, способных эффективно работать с документными таблицами произвольной структуры, представляет собой актуальную научно-техническую проблему, решение которой имеет важное значение для повышения эффективности процессов извлечения, интеграции и использования данных в различных прикладных областях.

Научная новизна. Основными результатами, составляющими научную новизну диссертационной работы, являются:

1. По сравнению с имеющимися формулировками АПТ, актуализированная структура совокупности задач АПТ является более релевантной за счет согласованной терминологии и многоуровневой декомпозиции.
2. В отличие от аналогов предлагаемый метод автоматизации распознавания таблиц базируется на использовании особенностей представления таблиц в НПОД. Впервые показано, как можно задействовать PDL-специфичную информацию для сегментации страниц и распознавания таблиц НПОД с целью улучшения качества их результатов.
3. В отличие от моделей документных таблиц, применяемых конкурентными решениями, созданная модель ориентирована на представление произвольной табличной структуры, что позволяет ей быть применимой к более широкому кругу сценариев АПТ.
4. В отличие от аналогов предлагаемый метод автоматизации анализа и интерпретации таблиц РК основан на пользовательском программировании правил. Впервые обеспечена поддержка произвольности структуры таблицы, а именно свободной компоновки, структурированности ячеек и иерархичности заголовков.
5. Впервые создан проблемно-ориентированный язык правил анализа и интерпретации документных таблиц. В отличие от языков общего назначения, он позволяет

сфокусироваться исключительно на реализации логики соответствующих этапов АПТ, упрощая тем самым разработку целевого программного обеспечения (ПО).

6. По сравнению с другими доступными инструментами АПТ разработанный комплекс инструментальных средств (КИС) в части распознавания таблиц НПОД дополнен анализом компоновки страниц и фильтрацией кандидатных случаев, а в части извлечения данных из таблиц РК отделяет правила их анализа и интерпретации от моделей представления и алгоритмов обработки.

Теоретическая ценность полученных результатов заключается в формировании теоретического базиса для решения задачи упрощения разработки ПО, предназначенного для извлечения данных из таблиц в документах неструктурированного формата (НПОД/РК) за счет поддержки произвольности табличной структуры. **Практическая значимость** подтверждается созданием технологии автоматизации процессов распознавания, анализа и интерпретации документных таблиц НПОД/РК. Разработанные инструментальные средства нашли применение в пяти научных и трех промышленных проектах 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.

Достоверность полученных результатов подтверждается представленными в диссертационной работе экспериментальными данными, качественным и количественным сравнением с имеющимися аналогами, а также программной реализацией. Разработанный КИС и материалы для воспроизведения проведенных экспериментов опубликованы в открытом доступе под свободными лицензиями.

Содержание и структура диссертации.

Во введении приводится общая характеристика диссертационной работы. Обосновывается актуальность выбранной темы, определяется научная новизна, теоретическая и практическая значимость основных результатов.

В первой главе содержит обзор проблематики АПТ и текущего состояния исследований в данной области. Приводится характеристика и структура комплекса ее задач, рассматриваются их постановки и существующие методы решения, а также доступные тестовые стенды и коллекции данных; очерчивается область применения существующих решений. Анализируются ограничения современных исследований и предлагаются перспективные направления для их развития.

Во второй главе предлагается метод автоматизации распознавания таблиц (РТ) в НПОД. Вводятся предварительные определения, составляющие модель представления страницы НПОД в процессе РТ. Рассматривается постановка задачи и предлагаемый метод ее решения. Подробно излагаются разработанные методики сегментации страницы документа, обнаружения и сегментации таблицы.

Третья глава посвящена методу автоматизации анализа и интерпретации таблиц (АИТ) РК на основе пользовательского программирования правил. Приводится постановка задачи и предлагаемый метод ее решения. Вводится модель представления физической и логической структуры документной таблицы. Описываются допустимые манипуляции с этой структурой в терминах данной модели, определяемые пользовательскими правилами. Подробно характеризуется проблемно-ориентированный язык правил для АИТ.

Четвертая глава посвящена разработке КИС на основе методов, предлагаемых в двух предыдущих главах. Рассматриваются архитектура и функциональность разработанного программного обеспечения для РТ и АИТ. Излагается методика настройки искусственных нейронных сетей для обнаружения таблиц. Приводятся примеры применения CRL-правил в различных часто встречающихся на практике сценариях. Представлены данные исследования поведения пользователей при работе с языком CRL.

Пятая глава демонстрирует результаты оценки производительности реализованных решений, а также их количественное и качественное сравнение с существующими аналогами. Первый раздел главы посвящен решениям для РТ в неразмеченных PDF-

документах на основе ПО TabbyPDF, а второй – решениям для АИТ регламентированной отчетности на базе ПО TabbyXL.

В заключении подведены итоги и обобщены результаты проведенных исследований.

Замечания по диссертации.

1. В автореферате описано, что CRL упрощает разработку для конечных пользователей, по представленному примеру (рис. 15) видно, что язык требует высокого уровня программирования. Насколько легко конечным пользователям, не являющимся программистами осваивать и продуктивно использовать CRL?

2. В автореферате подробно оценивается точность, но не описывается производительность и время обработки больших объемов данных. Какова вычислительная сложность предложенных методов и производительность инструментов при пакетной обработке тысяч документов.

3. В тексте автореферата присутствуют незначительные погрешности в виде опечаток и пунктуационных ошибок, не искажающие смысла изложения. Например, на странице 9 «выполнено при поддержки грантов» – правильно «при поддержке».

Отмеченные недостатки не влияют на положительную оценку всех научных и практических результатов диссертации.

Заключение. На основании всего изложенного считаю, что диссертация Шигарова Алексея Олеговича является самостоятельно выполненной научно-квалификационной работой, обладающей внутренним единством и содержащей новые научные результаты, имеющие важное значение для развития научно-методологического аппарата создания программного обеспечения для извлечения данных из документных таблиц. Диссертационная работа соответствует требованиям, предъявляемым п. 9 и п. 10 «Положения о присуждении ученых степеней» к докторским диссертациям, 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».

Доктор физико-математических наук,
заведующий кафедрой вычислительной
математики и кибернетики Северо-
Кавказского федерального университета

Бабенко Михаил Григорьевич

На включение персональных данных,
содержащихся в настоящем отзыве, в
документы, связанные с работой
диссертационного совета, и их дальнейшую
обработку согласен.

21.11.2021

Бабенко Михаил Григорьевич

Бабенко Михаил Григорьевич - доктор физико-математических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Доцент по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ». Заведующий кафедрой вычислительной математики и кибернетики факультета математики и компьютерных наук имени профессора Н.И. Червякова, ФГАОУ ВО «Северо-Кавказский федеральный университет». Адрес: 355009, г. Ставрополь, ул. Пушкина, 1. E-mail: mgbabenko@ncfu.ru. Тел: +7 906 440-02-19



ДПИСЬ УДОСТОВЕРЯЮ:

начальник отдела по
работе с сотрудниками УКА

И.С. ГОРБАЧЕВ