

ОТЗЫВ

**на автореферат диссертации на соискание ученой степени
доктора технических наук Шигарова Алексея Олеговича на тему:
«Методы и инструментальные средства автоматизации процессов
извлечения данных из таблиц электронных документов
неструктурированного формата»**

Диссертационная работа Шигарова Алексея Олеговича посвящена решению проблемы автоматизированного извлечения данных из неструктурированных таблиц.

На первый взгляд, данная проблема уже решена. В самом деле, как только таблица оформлена электронно, её содержание известно. Однако, на практике ситуация может быть совсем иной.

Предметом исследования в данной работе являются таблицы, представленные в виде растровых изображений, «закрытых» форматов, таких как PDF, а также в виде разнообразных Web-документов. Для решения задачи извлечения данных из множества подобных таблиц необходим общий подход и разработка прикладных решений, являющихся достаточно универсальными, и в то же время учитывающими особенности конкретных форм представления таблиц. Такой подход, соответствующая методика и инструментальные средства разработаны в диссертации Шигарова Алексея Олеговича, что придаёт его работе несомненную актуальность.

На основании содержания автореферата основными достоинствами диссертации являются следующие.

1. Предложенная в работе унифицированная структура задач автоматизированного понимания таблиц реализует междисциплинарную интеграцию, например, через сближение терминологии. Это упрощает взаимодействие между исследователями из разных областей. Системный подход, примененный автором, через декомпозицию способствует созданию модульных решений, где каждый этап решаемых задач может оптимизироваться независимо. Унифицированные решения задач автоматизированного понимания таблиц могут лечь в основу будущих стандартов в этой области.

2. Предложен инновационный метод конвертации таблиц из неструктурированных документов в редактируемый формат. Научная значимость метода в том, что он позволяет преодолеть ограничения методов, опирающихся только на визуальные признаки, например, линии сетки таблицы. Метод позволяет учитывать координаты и атрибуты таблиц (например, выравнивание текста) и улучшает сегментацию сложных таблиц, например, с объединенными ячейками. Метод также может адаптироваться к разным форматам, где доступна PDL-информация.

3. В рамках метода разработана модель таблиц. Определённая универсальность модели состоит в том, что в ней не ограничивается структура таблиц предопределенными шаблонами, в отличие от многих существующих моделей, предполагающих фиксированную компоновку. Модель позволяет сохранять контекст данных (например, иерархию

заголовков), что критично для задач семантического анализа и задач преобразования в RDF/онтологии.

4. В работе имеется впечатляющая практическая реализация полученных решений. Следует отметить проблемно-ориентированный язык правил анализа и интерпретации таблиц. Разработанный язык класса DSL упрощает написание правил за счёт семантики, адаптированной под табличные данные.

5. Результаты диссертации реализованы в программном обеспечении, построенном на основе современных вычислительных технологий и предоставленном в открытый доступ.

Замечания по автореферату.

1. В автореферате недостаточно полно освещены нейросетевые решения. В описании четвертой главы следовало поместить информацию о структуре применяемых нейронных сетей, характере обучающих и тестовых данных и результатах экспериментов.

2. В описании третьей главы проблемно-ориентированный язык правил представлен отдельными примерами команд. Формальная спецификация языка, не вся, но её важные элементы – позволили бы составить более полное представление о разработанном языке.

Общий вывод о работе.

Представленные в автореферате результаты формируют целостный подход к решению задач автоматизированного понимания таблиц, охватывая все ключевые элементы исследуемой проблемы. В работе имеются теоретическая новизна, практическая значимость и конкурентные преимущества, обусловленные отсутствием аналогов полученных решений.

Изложенная в автореферате работа соответствует требованиям ВАК к докторским диссертациям по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей», а ее автор, Шигаров Алексей Олегович, заслуживает присуждения ему ученой степени доктора наук.


Богатырев Михаил Юрьевич,
д.т.н., доцент, профессор кафедры
Информационной безопасности
ФГБОУ ВО «Тульский
государственный университет»
Адрес: 300045, г.Тула, ул.
Новомосковская, д.7, кв. 79
Email: okkambo@mail.ru
Телефон: +7 (920) 755-2479

Даю свое согласие на включение моих персональных данных в документы, связанные с работой диссертационного совета:

Подписи М.Ю. Богатырева заверяю:
Начальник Административно-кадрового управ

М.Ю. Богатырев

К.А. Дымус

