На правах рукописи

Шигаров Алексей Олегович

Технология извлечения табличной информации из электронных документов разных форматов

05.25.05 – Информационные системы и процессы,
правовые аспекты информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Иркутск – 2009

Работа выполнена в Учреждении Российской академии наук Институте динамики систем и теории управления Сибирского отделения РАН

Научный руководитель:	член-корреспондент РАН, доктор технических наук, Бычков Игорь Вячеславович
Официальные оппоненты:	доктор технических наук, Жижимов Олег Львович доктор технических наук, Тятюшкин Александр Иванович
Ведущая организация:	Государственное образовательное учреждение высшего профессионального образования «Иркутский государственный университет»

Защита состоится «5» февраля 2010 г. в 16:00 на заседании диссертационного совета ДМ 003.046.01 в Учреждении Российской академии наук Институте вычислительных технологий Сибирского отделения РАН по адресу: 630090, Новосибирск, пр. Академика Лаврентьева, 6

С диссертацией можно ознакомиться в специализированном читальном зале вычислительной математики и информатики ГПНТБ СО РАН

Автореферат разослан «30» декабря 2009 г.

Ученый секретарь
диссертационного совета,
доктор физико-математических наук,
профессор Чубаров Л. Б.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. Таблицы являются одним из наиболее используемых способов представления информации в документах. Часто такая табличная информация изначально представлена в базах данных. Однако базы данных не всегда доступны, открыты, содержат полную информацию. В связи с этим актуальность приобретают методы, системы и технологии извлечения табличной информации из документов, которые позволяют преобразовать её к требуемому виду, например, к отношениям в реляционных базах данных.

Актуальность данной проблематики подчеркивается в обзорах авторов Handley J.C. (1999), Lopresti D. и Nagy G. (1999, 2000), Zanibbi R. (2004), Embley D.W. (2006), e Silva A.C. (2006), Полевого Д.В. (2007). В литературе выделяется несколько задач связанных с извлечением табличной информации: 1) *обнаружение* — поиск на страницах документов областей, ограничивающих таблицы; 2) *сегментация* — разделение таблицы на отдельные ячейки, строки, столбцы; 3) *анализ функций ячеек* — определение того, какие функции выполняют отдельные ячейки таблицы (являются ли они заголовками или данными); 4) *структурный анализ* — определение связей между ячейками таблицы.

Несмотря на то, что в последние годы появились работы, в которых предлагаются некоторые методы и системы извлечения табличной информации из документов, нельзя считать эту проблему полностью решенной. Сложность автоматического извлечения табличных данных во многом обусловлена большим разнообразием форм изображения таблиц. Известные методы и системы преимущественно ориентированы на заранее определенные структуры и особенности таблиц, которые связаны со стандартами выбранной предметной области. От этого во многом зависит их эффективность. При этом они, как правило, решают только отдельные задачи, например, только обнаружение или сегментацию таблиц.

Автоматическое обнаружение и сегментация таблиц выполняется в некоторых системах оптического распознавания текста, например, «OmniPage» (Nuance Communications), «Cuneiform» (Cognitive Technologies), «FineReader» (ABBYY). Перечисленные системы ориентированы на «решёточную» структуру таблиц, характерную табличным процессорам, например, Excel. Это снижает их эффективность для обнаружения и сегментации таблиц со сложными структурами заголовков. Кроме того, оптическое распознавание символов выполняется с потерями информации. В системах «PDF2XL» (Cogniview) и «Solid Converter PDF» (Solid Documents), в частности, выполняется преобразование таблиц из документов PDF в файлы Excel/Word. Эти системы также ориентированы на «решёточную» структуру таблиц и выполняют только их обнаружение и сегментацию.

В статистических отчетах (государственных, медицинских, финансовых) основная информация представлена в виде так называемых *статистических таблиц*[1]. Вне зависимости от национальной или корпоративной принадлежности такие таблицы обладают достаточно схожей структурой. Большинство таких отчетов доступно в электронном виде, где таблицы, как правило, являются *машиночитаемым текстом*, т. е. электронным текстом, который хранится в виде строк символов. Однако в литературе не представлены методы или системы извлечения табличной информации, которые с одной стороны являются комплексными, т. е. выполняют обнаружение, сегментацию, анализ функций ячеек и структурный анализ таблиц, а с другой стороны ориентированы на структуру и особенности статистических таблиц, в частности, публикуемых Росстатом. Таким образом, разработка комплексной технологии извлечения табличной информации, которая ориентирована на структуру и особенности статистических таблиц, представленных в виде машиночитаемого текста в электронных документах, является актуальной задачей.

Цель диссертационной работы состоит в создании технологии извле-

[1] «Большая советская энциклопедия. Изд. 3-е» – М.: Советская энциклопедия. Т.25 «Струнино-Тихорецк». 1976. С. 161-162. http://slovari.yandex.ru/dict/bse/article/00077/08800.htm

чения табличной информации из электронных документов разных форматов, которая автоматизирует обнаружение, сегментацию, анализ функций ячеек и структурный анализ статистических таблиц.

Основные задачи диссертационной работы.

1. Анализ представления статистических таблиц в документах.

2. Разработка моделей страницы документа и таблицы, предназначенных для представления данных в процессе извлечения табличной информации из электронных документов, на основе проведенного анализа.

3. Разработка методов автоматического обнаружения, сегментации, анализа функций ячеек и структурного анализа статистических таблиц на основе предложенных моделей.

4. Разработка информационной системы извлечения табличной информации из электронных документов на основе предложенных методов.

5. Проверка созданной технологии на задачах автоматизации ввода больших объемов табличной информации из электронных статистических отчетов в базы данных.

Методы исследования: теория множеств, теория баз данных, методы машинной графики, объектно-ориентированное программирование.

Научная новизна. Впервые предложена технология извлечения табличной информации, представленной в виде машиночитаемого текста в электронных документах разных форматов, которая ориентирована на структуру и особенности статистических таблиц и является комплексной, т. е. выполняет их обнаружение, сегментацию, анализ функций ячеек и структурный анализ.

Практическая значимость. Результаты диссертационной работы могут использоваться в задачах извлечения информации и управления данными. В частности, предлагаемая технология может использоваться для автоматизации ввода в базы данных информации из статистических таблиц,

представленных в виде машиночитаемого текста в электронных документах разных форматов. При этом данная технология позволяет снизить затраты и повысить качество формирования баз данных. Работа выполнена при поддержке РФФИ, грант 09-07-12017-офи_м.

Внедрение. Результаты диссертационной работы успешно использовались в Министерстве сельского хозяйства Иркутской области для ввода информации из электронных статистических отчетов Территориального органа федеральной службы государственной статистики по Иркутской области (Иркутскстата) в базу данных (БД) автоматизированной информационной системы (АИС) «Каскад». Предлагаемая технология внедрена в Институте систем энергетики им. Л.А. Мелентьева СО РАН, где используется при создании хранилища данных в составе информационной инфраструктуры исследований в энергетике.

Защищаемые положения.

1. Модель страницы документа, которая служит для представления данных страницы, используемых в процессе извлечения табличной информации.

2. Модель структурного описания таблицы, которая предназначена для представления табличных заголовков и данных, а также связей между ними.

3. Методы обнаружения, сегментации, анализа функций ячеек и структурного анализа статистических таблиц, которые обеспечивают извлечение и структурирование табличной информации, содержащейся в электронных документах.

Личный вклад автора. Основные результаты диссертационной работы получены автором лично, а именно: предложены модель страницы документа, эвристические методы обнаружения, сегментации, анализа функций

ячеек и структурного анализа статистических таблиц; разработана информационная система для извлечения табличной информации из метафайлов EMF (Enhanced Metafiles); создана технология извлечения табличной информации из электронных документов разных форматов. В неделимом соавторстве с А.Е. Хмельновым получена модель структурного описания таблицы. В неделимом соавторстве с А.Е. Хмельновым, И.В. Бычковым и Г.М. Ружниковым получено применение предлагаемой технологии для автоматизации ввода статистической информации в базу данных АИС «Каскад». В работах [2–4, 7, 8] автором лично предложен эвристический метод обнаружения таблиц и технология извлечения табличной информации из электронных документов разных форматов. В работах [5, 10–12] автором в неделимом соавторстве с А.Е. Хмельновым предложена модель структурного описания таблицы.

Представление работы. Основные результаты работы докладывались на научно-практических конференциях: Международной конференции «Математические и информационные технологии» (Будва, Черногория, 2009 г.); IX международной конференции «Распознавание образов и анализ изображений: новые информационные технологии» (Нижний Новгород, 2008 г.); XII, XIII и XIV всероссийской конференции «Информационные и математические технологии в науке и управлении» (Иркутск, 2007, 2008, 2009 гг.); Всероссийской конференции «Математическое моделирование и вычислительно-информационные технологии в междисциплинарных научных исследованиях» (Иркутск, 2009 г.); VI и IX школе-семинаре «Математическое моделирование и информационные технологии» (Иркутск, 2005, 2007 гг.); Школе-семинаре молодых ученых «Информационные технологии и моделирование социальных эколого-экономических систем» (Иркутск, 2008 г.); семинаре «Ляпуновские чтения и презентация информационных технологий» (Иркутск, 2007, 2008, 2009 гг.).

Публикации. По теме диссертации опубликовано 12 научных работ [1–12], в т. ч. 3 публикации [1–3] в изданиях, рекомендованных ВАК. Получено

4 свидетельства об официальной регистрации программ для ЭВМ в Роспатенте: №№ 2008614328, 2008614330, 2008614331, 2008614332 (2008 г.).

Структура и объем работы. Диссертация состоит из введения, 4-х глав, заключения, списка литературы, включающего 103 источника, и 4-х приложений. Основное содержание диссертации изложено на 132 страницах текста, общее количество страниц — 141.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении приводится общая характеристика работы.

В главе 1 приводится анализ процесса извлечения табличной информации из документов, рассматриваются известные методы и системы извлечения табличной информации.

Рассматривается разнообразие используемых в документах таблиц, а также структура и особенности статистических таблиц, Рис. 1. Обсуждаются форматы входных данных, которые используются в известных методах и системах извлечения табличной информации. Показано, что в основном в качестве входных данных применяются либо ASCII-текст без графического форматирования (не поддерживает всех возможностей современных текстовых и табличных процессоров), либо растровые изображения документов (требуют оптического распознавания текста), либо Web-страницы формата HTML (таблицы используются для компоновки Web-страниц).

В диссертации предлагается использовать в качестве входных данных метафайлы. Это позволяет извлекать табличную информацию, представленную в виде машиночитаемого текста в электронных документах разных форматов, например, DOC, XLS, PDF (с латиницей), HTML, ASCII-текст. Поскольку электронные документы таких форматов могут преобразовываться в метафайлы посредством виртуальной печати. При этом машиночитаемый текст исходных документов остается в метафайлах машиночитаемым. Следует отметить, что в отличие от файлов форматов PostScript и PDF метафайлы могут интерпретироваться с помощью GDI (Graphics Device Interface, части

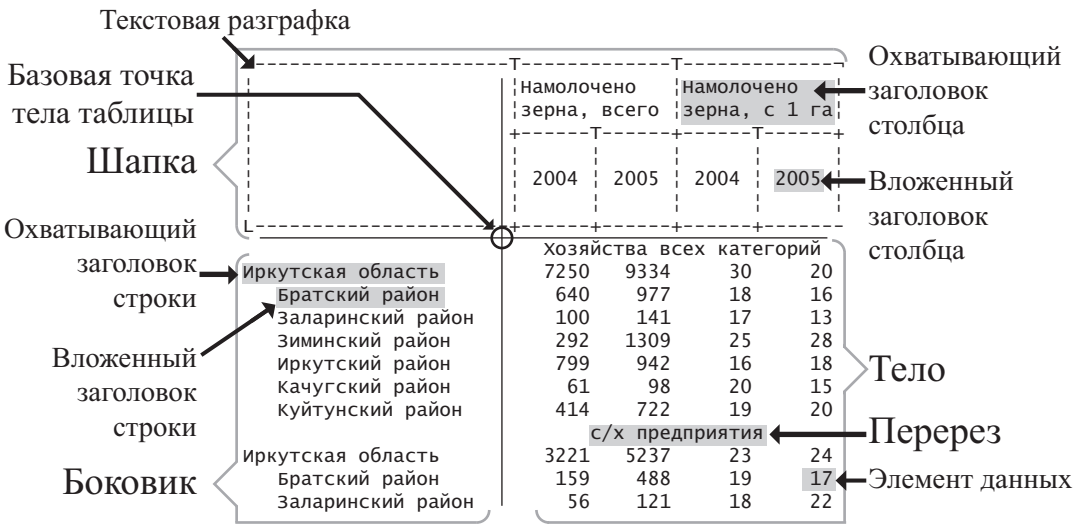

Рис. 1. Пример статистической таблицы

Windows API). Также рассматриваются известные методы и системы извлечения табличной информации из документов. Обсуждаются их ограничения. На основе проведенного анализа предлагается технология извлечения табличной информации из электронных документов, основные компоненты и этапы которой приводятся на Рис. 2.

В главе 2 рассматривается обработка страниц документов, предлагается оригинальный эвристический метод обнаружения таблиц.

Описываются особенности и ограничения обрабатываемых таблиц, Рис. 1. Предлагается теоретико-множественная модель страницы документа, которая служит для представления данных обрабатываемой страницы. Основными объектами этой модели являются линейки (линии разграфки), текстовые элементы, текстовые блоки, строки, табличные регионы, табличные области, Рис. 3. Эти объекты формируются снизу вверх, Рис. 4. В предлагаемых методах используется анализ промежутков пустого места на странице (т. е. места, не занятого текстовыми блоками). Для этого предлагается алгоритм сегментации пустого места и выделение среди полученных сегментов вертикальных и горизонтальных промежутков, Рис. 5.

Описывается обработка и интерпретация метафайлов с помощью GDI, а также формирование из записей метафайлов текстовых элементов и лине-

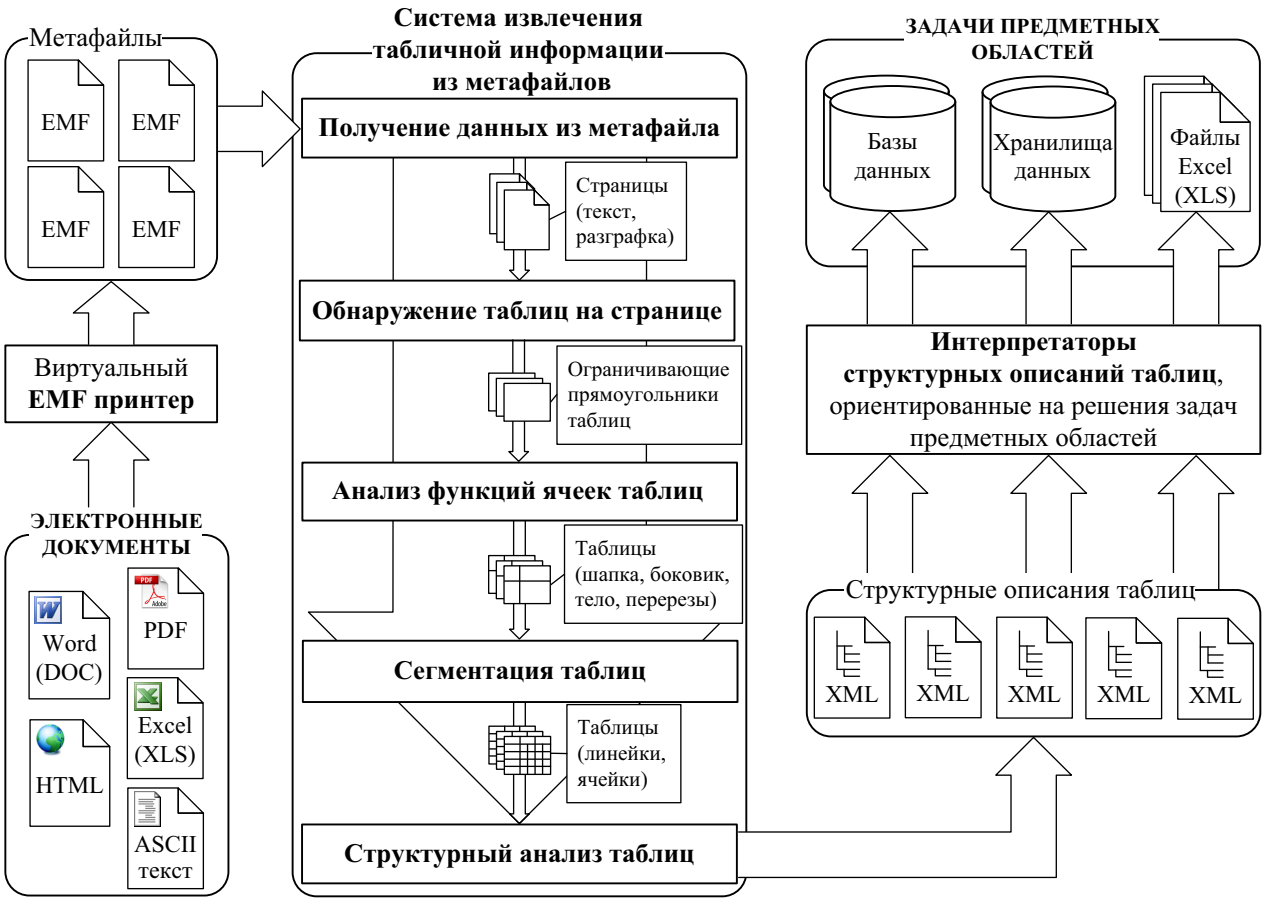

Рис. 2. Технология извлечения табличной информации из электронных документов

ек. Предполагается, что каждый метафайл представляет отдельную страницу. Рассматривается предобработка страницы, которая, в частности, предусматривает исключение из текста текстовой разграфки (линеек, образованных символами псевдографики). Линейки текстовой разграфки преобразуются к графическим линейкам.

Предлагается метод обнаружения таблиц на странице, т. е. поиска ограничивающих прямоугольников таблиц — табличных областей. Для этого текстовые элементы, близко расположенные в одной строке текста друг к другу и при этом не разделенные линейками, объединяются в текстовые блоки, Рис. 6. Близость расположения двух текстовых элементов вычисляется с помощью их шрифтовых метрик. Для текстовых блоков вычисляются ограничивающие прямоугольники по вложенным в них текстовым элементам. Текстовые блоки группируются в строки. При этом если у двух текстовых бло-

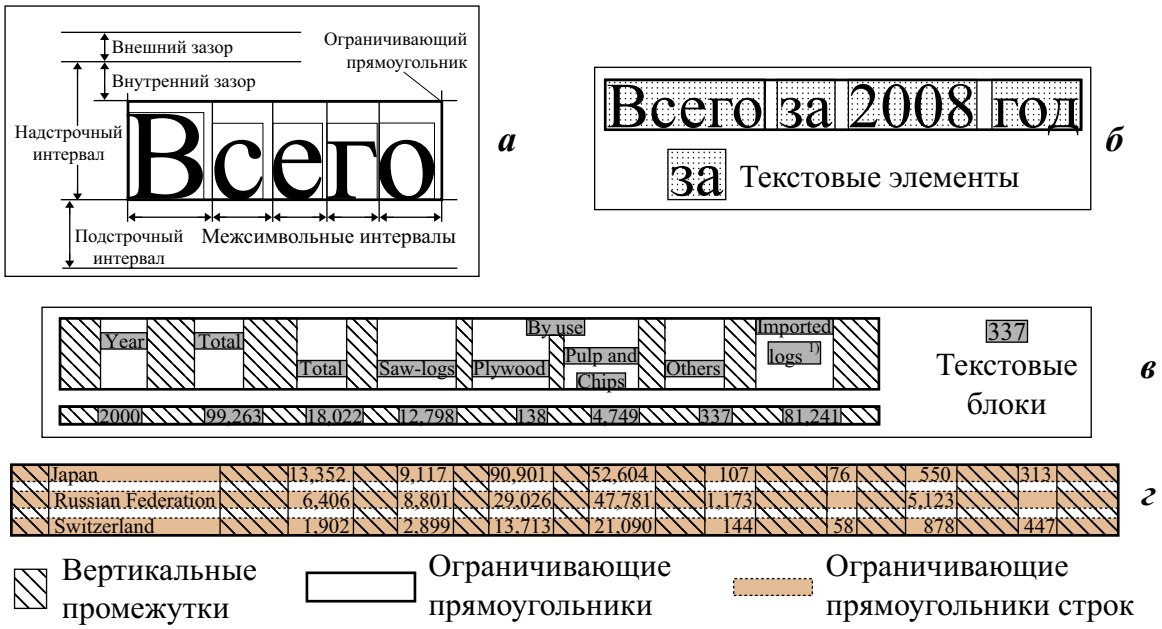

Рис. 3. Основные объекты страницы: текстовый элемент ($а$), текстовый блок ($б$), строки ($в$), табличный регион ($г$)

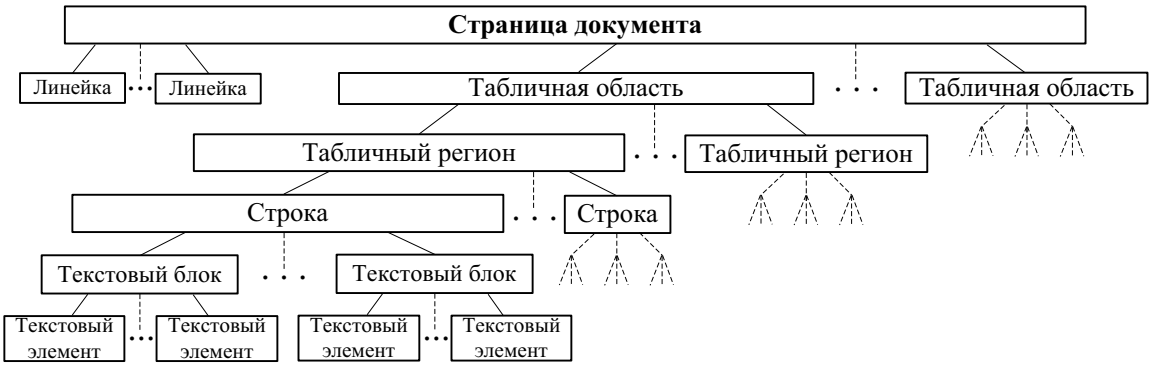

Рис. 4. Порядок формирования объектов страницы документа снизу вверх

ков их проекции на ось Y пересекаются, то они принадлежат одной строке. На странице среди всех строк выбираются строки табличного вида. Для этого используется ряд эвристик о составе строк табличного вида. Например, такая строка должна охватывать не менее двух текстовых блоков и иметь ширину пустого места относительно всей своей ширины не менее заданного порога. На странице выполняется поиск последовательностей подряд расположенных сверху вниз строк табличного вида, которые имеют схожее расположение проекций на ось X своих вертикальных промежутков. Каждая такая

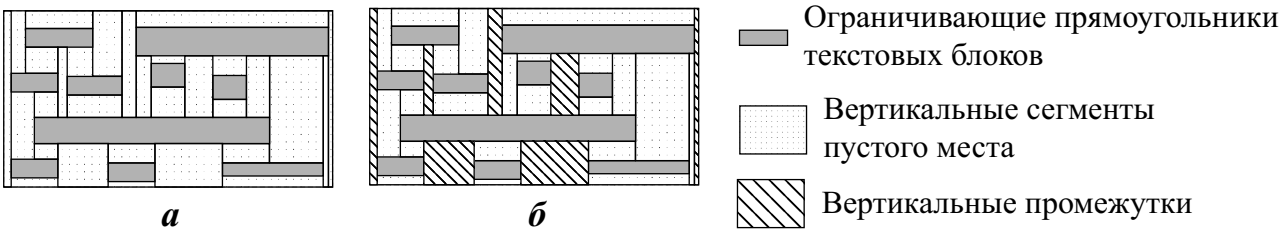

Рис. 5. Сегментация пустого места ($а$) и выделение вертикальных промежутков ($б$)

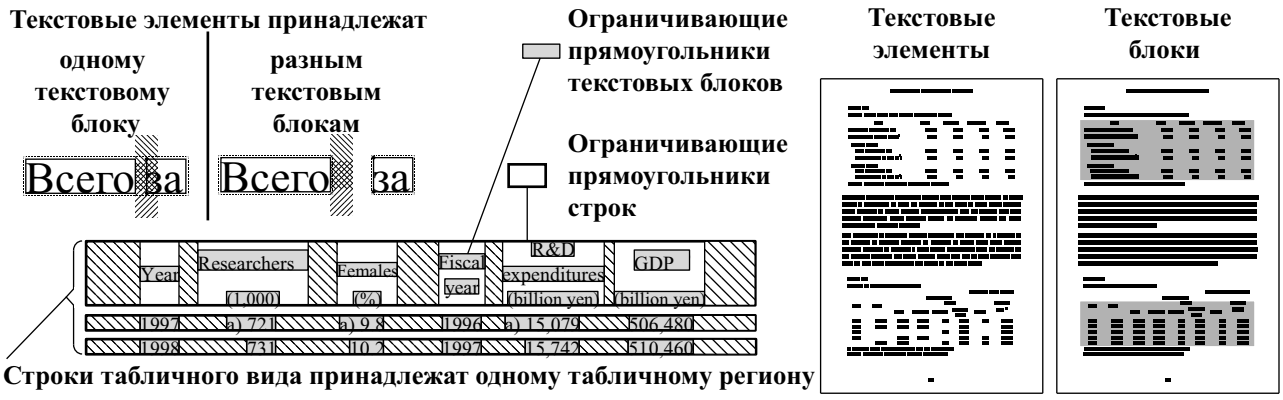

Рис. 6. Обнаружение таблиц на странице документа

последовательность образует отдельный табличный регион. Строки страницы проходятся сверху вниз, если найден табличный регион, то его строки исключаются из дальнейшего поиска. Затем из соседних табличных регионов, которые схожи по расположению проекций на ось X своих вертикальных промежутков, формируются табличные области. Предполагается, что каждая табличная область ограничивает таблицу.

В главе 3 описывается анализ и обработка таблиц, предлагаются оригинальные эвристические методы анализа функций ячеек, сегментации и структурного анализа таблицы.

Предлагается теоретико-множественная модель таблицы, которая представляет объекты, связанные с обрабатываемой таблицей. Формализованы ячейка и таблица, составленная из наборов текстовых блоков, строк, вертикальных и горизонтальных линеек, ячеек, базовой точки тела и ограничивающего прямоугольника. Обсуждается предобработка входных данных — табличной области и содержащихся внутри неё текстовых блоков и линеек,

Рис. 7. Поиск базовой точки тела таблицы

по которым выполняется первоначальное формирование таблицы.

Описывается анализ функций ячеек таблицы. Функция (роль) ячейки зависит от её расположения относительно базовой точки тела таблицы. Эта точка делит таблицу на шапку, боковик и тело. Предлагаемый метод анализа функций ячеек строится, как поиск базовой точки табличного тела, Рис. 7. Для этого внутри таблицы определяется область поиска этой точки, которая начинается непосредственно под самым нижним охватывающим заголовком столбца. Эта область имеет «решёточную» структуру ячеек. Она сегментируется на отдельные ячейки с помощью вертикальных промежутков и ограничивающих прямоугольников строк таблицы. Данные, содержащиеся в теле статистической таблицы, являются числами или специальными обозначениями из ограниченного набора. С помощью заранее заданных регулярных выражений каждой непустой ячейке по её тексту сопоставляется один из следующих типов данных: «числа», «даты» или остальной «текст». По ячейкам, содержащим «числа», строится ограничивающий прямоугольник тела таблицы. Вершина в левом верхнем углу этого прямоугольника является базовой точкой тела таблицы. Если непосредственно над этой точкой располагаются табличные строки, включающие по одному текстовому блоку, то её y-координата корректируется с помощью эвристик о расположении перерезов и заголовков. Кроме того, отдельно выделяются строки таблицы, содержащие перерезы.

Рассматривается сегментация таблицы. Статистические таблицы, как правило, имеют только частичную разграфку или не имеют её вовсе. Предла-

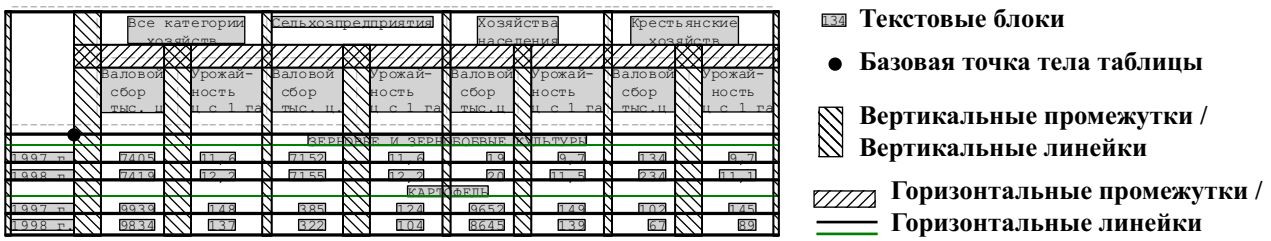

Рис. 8. Восстановление полной разграфки таблицы

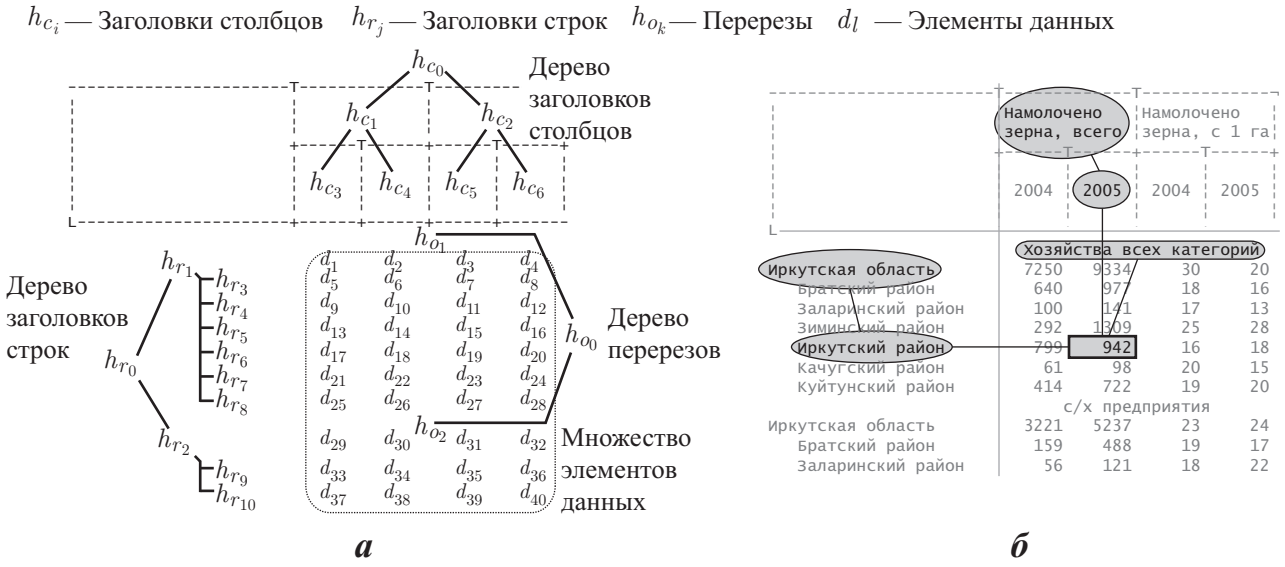

Рис. 9. Компоненты структурного описания таблицы (a) и пример связывания элемента данных с заголовками ($б$)

гаемая сегментация таблицы выполняется, как восстановление полной разграфки таблицы, Рис. 8. По вертикальным промежуткам таблицы восстанавливаются вертикальные линейки. Далее, по горизонтальным промежуткам таблицы восстанавливаются горизонтальные линейки внутри её шапки. Горизонтальные линейки под шапкой восстанавливаются по ограничивающим прямоугольникам табличных строк. Восстановленные линейки составляют полную табличную разграфку, которая корректируется с помощью исходной табличной разграфки при её наличии. С помощью своей полной разграфки таблица сегментируется на отдельные ячейки.

Предлагается теоретико-множественная модель структурного описания таблицы, которая служит для представления содержимого заголовков, эле-

ментов данных и связей между ними. Предлагаемое структурное описание таблицы включает в себя дерево заголовков столбцов, дерево заголовков строк, дерево перерезов и множество элементов данных, Рис. 9, *а*. Эти деревья соответствуют тем иерархиям вложенности, которые образуют заголовки (несмотря на то, что перерезы не имеют вложенности, их также удобнее рассматривать, как дерево). Вложенные заголовки являются в этих деревьях подузлами охватывающих заголовков. Корнями этих деревьев являются пустые узлы. Каждый элемент данных сопоставляется с одним заголовком столбца, одним заголовком строки и одним перерезом, Рис. 9, *б*.

Предлагается метод структурного анализа таблицы для формирования её структурного описания. Выполняется эвристический анализ компоновки ячеек таблицы. Из содержимого ячеек формируются: 1) дерево заголовков столбцов, 2) дерево заголовков строк, 3) дерево перерезов и 4) множество элементов данных. Выполняется связывание элементов данных с заголовками.

Предложенные методы обнаружения, сегментации, анализа функций ячеек таблицы и структурного анализа статистических таблиц обеспечивают их комплексное извлечение из электронных документов.

В главе 4 рассматривается применение созданной технологии.

Предлагается информационная система извлечения табличной информации из метафайлов, которая реализует предлагаемые методы. Эта система имеет графический пользовательский интерфейс, который визуализирует процесс извлечения табличной информации. На выбранной странице этот процесс выполняется поэтапно: 1) обнаружение, 2) анализ функций ячеек, 3) сегментация и 4) структурный анализ таблиц. При этом пользователь при необходимости может вручную корректировать результаты каждого из этих этапов. Данная система позволяет представить получаемые структурные описания таблиц в виде XML, Рис. 10, структура которого описана на языке XML Schema.

```
<table name="Таблица 1">
  <columnHeader text="Заголовки столбцов" id="0">
    <columnHeader text="Намолочено зерна, всего" id="14581672">
      <columnHeader text="2004" id="14581896"/> [...]
    </columnHeader> [...]
  </columnHeader>                                   Дерево заголовков столбцов
  <rowHeader text="Заголовки строк" id="0">
    <rowHeader text="Иркутская область" id="14582344">
      <rowHeader text="Братский район" id="14582400"/> [...]
    </rowHeader> [...]
  </rowHeader>                                      Дерево заголовков строк
  <cutinHeader text="Перерезы" id="0">
    <cutinHeader text="Хозяйства всех категорий" id="14582848"/> [...]
  </cutinHeader>                                    Дерево перерезов
  <data>
    <dataElement text="7250" colId="14581896" rowId="14582344" cutId="14582848"/>
    <dataElement text="640" colId="14581896" rowId="14582400" cutId="14582848"/>
    [...]
  </data>                                           Множество элементов данных
</table>
```

Рис. 10. Фрагмент XML представления структурного описания таблицы

Таблица 1. Экспериментальная оценка

Обнаружение:	таблиц	базовых точек тел таблиц	линеек
Точность	84,5%	91,4%	86,2%
Полнота	91,7%	X	82,5%

Приводится экспериментальная оценка данной системы, Таблица 1. Используется две оценки: 1) *точность* — процент количества корректно обнаруженных таблиц/базовых точек тел таблиц/линеек к общему количеству обнаруженных соответственно таблиц/базовых точек тел таблиц/линеек; 2) *полнота* — процент количества корректно обнаруженных таблиц/линеек к общему числу существующих соответственно таблиц/линеек. Экспериментальные данные были составлены из государственных статистических отчетов России, США, Евросоюза, Японии, а также из финансовых отчетов различных компаний. Они были представлены в форматах: PDF, DOC, XLS, HTML. Всего для оценки эффективности обнаружения таблиц/базовых точек тел таблиц было обработано 425 страниц, содержащих 518 таблиц. Для оценки эффективности обнаружения линеек из экспериментальных данных случайным образом было выбрано 44 страницы, содержащих 51 таблицу с 275 вертикальными и 1046 горизонтальными линейками.

Описывается автоматизация ввода статистической информации в БД АИС «Каскад» с помощью предлагаемой технологии. Неполнота представления статистической информации в базах данных Иркутскстата не позволяет организовать прямое преобразование необходимых данных в АИС «Каскад». Поэтому публикуемые электронные статистические отчеты Росстата являются основным источником необходимых данных. Эти отчеты представлены в форматах DOC, XLS, plain-text. При этом большинство таблиц в отчетах формата DOC являются включениями ASCII-текста, остальные являются табличными объектами Word. Каждое структурное описание извлеченной таблицы преобразуется в промежуточное представление, которое состоит из 1) таблицы формата СУБД «Paradox» (хранит в реляционном виде данные из статистической таблицы) и 2) текстового FNI (Field Name Information) файла (хранит информацию о структуре табличных заголовков и их связях с полями реляционного отношения). Для этого деревья заголовков структурного описания таблицы объединяются в одно дерево показателей. С помощью регулярных выражений в дереве показателей идентифицируются заголовки, обозначающие лексически «время» и «территории». Эти заголовки исключаются из дерева показателей и образуют два измерения — «время» и «территории». Также из дерева показателей исключаются игнорируемые заголовки, указывающие на вычислимые данные. Формируется реляционное отношение: элементы данных связанные с одним листом дерева показателей образуют поле, также два поля образуют соответственно значения измерений «время» и «территории». Формируется FNI файл, в котором каждой метке поля сопоставляется путь из дерева показателей. В БД АИС «Каскад» информация организована в виде дерева, узлами которого являются показатели из статистических отчетов. Для каждого промежуточного представления выполняется связывание со структурой БД АИС «Каскад», далее осуществляется автоматический ввод его данных.

Применение предлагаемой технологии для наполнения БД АИС «Кас-

кад» позволило снизить затраты и повысить качество при вводе в неё информации из электронных статистических отчетов.

В заключении приводятся основные полученные результаты диссертационной работы, обсуждаются перспективные направления их развития.

Основные полученные результаты.

1. Разработана модель страницы документа, которая служит для представления данных страницы, используемых в процессе извлечения табличной информации из электронных документов.

2. Разработана модель структурного описания таблицы, которая предназначена для представления заголовков и данных таблицы, а также связей между ними.

3. Разработаны методы обнаружения, сегментации, анализа функций ячеек и структурного анализа таблиц, ориентированные на структуру и особенности статистических таблиц.

Список публикаций

[1] Шигаров А.О. Технология извлечения табличной информации из электронных документов разных форматов [Текст] / Шигаров А.О. // Современные технологии. Системный анализ. Моделирование. – 2009. – № 3 (23). – С. 97–102.

[2] Бычков И.В. Эвристический метод обнаружения таблиц в разноформатных документах [Текст] / Бычков И.В., Ружников Г.М., Хмельнов А.Е., Шигаров А.О. // Вычислительные технологии. – 2009. – Т. 14, № 2. – С. 58–73.

[3] Shigarov A.O. A method for table detection in metafiles [Текст] / Shigarov A.O., Bychkov I.V., Khmel'nov A.E., Ruzhnikov G.M. // Pattern Recognition and Image Analysis. – 2009. – Vol. 19, No 4. P. 693–697.

[4] Бычков И.В. Метод обнаружения таблиц в метафайлах [Текст] / Бычков И.В., Ружников Г.М., Хмельнов А.Е., Шигаров А.О. // Современные технологии. Системный анализ. Моделирование. – 2008. – Спецвыпуск. – С. 47–51.

[5] Хмельнов А.Е. Метод извлечения таблиц из неформатированного текста [Текст] / Хмельнов А.Е., Шигаров А.О. // Вычислительные технологии. – 2008. – Т. 13, Спец. выпуск 1. – С. 93–101.

[6] Шигаров А.О. Автоматизированная система извлечения табличной информации из метафайлов [Текст] / Шигаров А.О. // Труды XIV Всероссийской конференции «Информационные и математические технологии в науке и управлении». – Иркутск, 2009. – Т. 2. - С. 218–224.

[7] Bychkov I.V. A method for table detection in metafiles [Текст] / Bychkov I.V., Hmelnov A.E., Ruzhnikov G.M., Shigarov A.O. // In Proc. 9th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies. – Nizhni Novgorod, 2008. – Vol. 1. – P. 66–69.

[8] Хмельнов А.Е. Сегментация страницы документа для обнаружения таблиц [Текст] / Хмельнов А.Е., Шигаров А.О. // Труды XIII Всероссийской конференции Информационные и математические технологии в науке и управлении. – Иркутск, 2008. – Ч. 2. – С. 244–251.

[9] Шигаров А.О. Метод обнаружения таблиц в метафайлах [Текст] / Шигаров А.О. // Материалы Школы-семинара молодых ученых Информационные технологии и моделирование социальных эколого-экономических систем. – Иркутск, 2008. – С. 58–61.

[10] Хмельнов А.Е. Метод извлечения статистических таблиц из неформатированного текста [Текст] / Хмельнов А.Е., Шигаров А.О. // Труды XII Всероссийской конференции Информационные и математические технологии в науке и управлении. – Иркутск, 2007. – Ч. 2. – С. 91–99.

[11] Хмельнов А.Е. Извлечение таблиц из неформатированного текста [Текст] / Хмельнов А.Е., Шигаров А.О. // Доклады 13–й Всероссийской конференции Математические методы распознавания образов (ММРО–13). – Зеленогорск, 2007. – С. 551–553.

[12] Хмельнов А.Е. Извлечение статистических таблиц из неформатированного текста [Текст] / Хмельнов А.Е., Шигаров А.О. // Материалы IX Школы-семинара Математическое моделирование и информационные технологии. – Иркутск, 2007. – С. 167–169.

Редакционно-издательский отдел
Учреждения Российской академии наук
Института динамики систем и теории управления
Сибирского отделения РАН
664033, Иркутск, ул. Лермонтова, 134
Подписано в печать 28.12.2009
Формат бумаги 60 x 84 1/16, объем 1,25 п.л.
Заказ № 10. Тираж 100 экз.
Отпечатано в ИДСТУ СО РАН