Автоматизированное понимание таблиц на основе системы исполнения правил

© А.О. Шигаров
Институт динамики систем и теории управления СО РАН
Иркутск
shigarov@icc.ru

Аннотация

В работе обсуждаются вопросы автоматизации процесса понимания таблиц, т.е. восстановления изначально отсутствующей в них информации о семантических отношениях (пары вида, ячейка-роль, метка-значение, метка-метка, метка-измерение). Предлагается подход, при котором понимание таблицы реализуется как исполнение правил анализа табличной структуры. На основе этого подхода разработана система для массового преобразования неструктурированной табличной информации, представленной в формате табличного процессора Excel, к структурированному виду. Результатом понимания таблиц являются структурированные данные — таблицы в канонической форме, которые структурно соответствуют таблицам реляционной базы данных. Полученные экспериментальные результаты показывают эффективность применения предлагаемого подхода для широкого класса сводных таблиц из статистических отчетов.

1 Введение

По оценки исследователей Merrill Lynch [16] примерно 80 процентов всей бизнес информации представлено в неструктурированном виде. Такая информация не имеет предопределенной формальной модели данных (например, научная статья, финансовый отчет, сообщение электронной почты) [1] и является противоположностью структурированной информации (например, реляционным базам данных).

Многие исследователи, в том числе, W. Inmon [11-12], отмечают важность вопросов интеграции неструктурированной информации. Одним из наиболее интересных вопросов является интеграция неструктурированных текстов, включая таблицы. Многие слабоструктурированные (ASCII-текст, файлы печати PDF и др.) и полуструктурированные (документы Word, книги Excel, HTML страницы и др.) документы [7] содержат таблицы. Такие таблицы главным образом адресованы для восприятия человеком. Они не предназначены напрямую для высокоуровневой машинной обработки, например, выполнения запросов к данным по аналогии с SQL (Structured Query Language). Поэтому они также являются примером неструктурированной информации.

На практике решения многих задач связаны с необходимостью извлекать информацию из таких таблиц и загружать её в базы данных. Поскольку, таблицы, представленные в неструктурированном виде, часто оказываются единственным доступным источником информации. Только после преобразование такой табличной информации к структурированной форме она становится доступной для использования в бизнес-аналитике, включая, аналитическую обработку в реальном времени (OLAP), интеллектуальный анализ данных, и извлечение знаний.

В литературе рассматриваются следующие задачи, которые являются преобразованием неструктурированной табличной информации к структурированному виду.

1) Каноникализация таблицы [2, 19] — приведение её к канонической форме, которая структурно соответствует таблице реляционной базы данных.

2) Извлечение информации из таблицы [5] является аналогом задачи извлечения информации из текста и состоит в выборочном извлечении фактов, формирующих целевую базу данных.

3) Понимание таблицы [5, 9] состоит в восстановлении отношений между метками (заголовками) и значениями данных, а также между метками и измерениями (доменами).

Как определяется в работе [9] понимание таблиц в общем случае включает следующие этапы: (1) обнаружение таблицы (поиск позиций ограничивающего прямоугольника таблицы внутри источника); (2) распознавание таблицы (разделение её на отдельные ячейки); (3) функциональный

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

анализ (определение того, какую роль играет ячейка в таблице); (4) структурный анализ (определение связей между ячейками); и (5) интерпретацию таблицы (извлечение фактов из таблицы). В настоящей работе обсуждается автоматизация следующих из перечисленных этапов понимании таблиц: (3) функционального и (4) структурного анализ, и (5) интерпретации таблицы.

2 Родственные работы

Существует огромное разнообразие способов изображения таблиц. Это приводит к высокой сложности анализа и обработки неструктурированной табличной информации. Как показано в обзорах [4, 5, 13, 22], посвященных проблемам анализа и обработки таблиц, сейчас наиболее изучены, хотя и не решены полностью, проблемы обнаружения и распознавания таблиц. При этом проблемы высокоуровневого анализа и интерпретации таблиц остаются менее изученными.

Вопросы понимания таблиц, связанные с задачами их (3) функционального и (4) структурного анализ, а также (5) интерпретации, рассматриваются в ряде работ [2, 4, 6, 8, 10, 19, 23–24]. Ниже приводится анализ некоторых из них.

В работах Douglas S. и др. [2] и Tijerino Y. и др. [19] рассматривается преобразование (структурирование) табличной информации, называемое каноникализацией таблицы. В работе Douglas S. и др. предлагается метод интерпретации и каноникализации таблиц, которые содержатся в спецификациях, используемых в строительной промышленности. Для этого они предлагают использовать обработку естественного языка на основе онтологии предметной области (подъязыка спецификаций строительной промышленности).

Предлагаемый Tijerino Y. и др. [19] способ каноникализации основан на использовании библиотеки фреймов, содержащей знания о лексическом содержании таблиц. Каждый фрейм данных описывает один тип данных и используется для отнесения выражений на естественном языке (табличных заголовков и значений) к этому типу. Для описания типов данных ими предлагается использовать регулярные выражения, словари и некоторые открытые ресурсы, например, WordNet [21].

В перечисленных работах [2–19] предлагаются методы каноникализации таблиц, основанные на анализе и интерпретации представленной в таблицах естественно-языковой информации. На практике этого не всегда достаточно, для более точного и полного извлечения информации из таблицы часто также требуется анализ пространственной и графической информации.

W. Gatterbauer и др. в работе [8] напротив предлагают предметно-независимый метод извлечения информации из HTML таблиц, основанный на анализе исключительно пространственной и стилевой информации в формате CSS2 (Cascading Style Sheets Level 2). В частности, ими предлагается выполнять интерпретацию таблиц (восстановление семантических отношений) на основе эвристик о стилевой информации подготовленного для набора наиболее общих типов изображения web-таблиц.

В работе D.W. Embley и др. [6] предлагаются методы обнаружения таблиц внутри HTML страниц, и извлечения из них информации. При этом предполагается, что таблица может включать вложенные таблицы на связанных страницах. В частности, для поиска атрибутов (меток) и значений (данных) среди содержания ячеек таблицы предлагается использовать онтологии, специально разрабатываемые для извлечения данных. Такие онтологии извлечения помимо понятий (объектов), отношений и ограничений содержат привязанные к объектам фреймы, которые с помощью регулярных выражений позволяют связать содержание таблицы с объектами онтологии. Для связывания атрибутов со значениями, дополнительно к онтологиям извлечения используется набор эвристик о пространственной структуре и содержании таблиц.

В отличие от приведенных исследований нами предлагается автоматизировать понимание таблиц за счет анализа и интерпретации, как их естественно-языковой, так и пространственной и графической (стилевой) информации.

3 Представление фактов о таблицах

Для понимания таблиц нами предлагается подход, основанный на исполнении правил анализа структуры таблиц. Идея, лежащая в основе предлагаемого подхода, состоит в следующем. Обычно внутри тематической коллекции документов от одного поставщика таблицы компонуются и форматируются однообразно. Для такой коллекции документов можно определить набор формализованных правил анализа табличной структуры, который удовлетворяет всем или почти всем ее таблицам. Эти правила можно представить в виде базы знаний, а процесс восстановления семантических отношений в таблице реализовать как логический вывод. При этом база фактов, используемая в процессе логического вывода, может включать информацию о пространственном, графическом и естественно-языковом содержании таблицы.

3.1 Базовые предположения о таблицах

На основе ограничений табличной структуры, характерных для представлений табличной информации в широко распространенных форматах данных, таких как Excel, Word, HTML и LaTeX, предлагается достаточно общая модель таблицы CELLS, которая ориентирована на представление фактов о табличной информации в процессе логического вывода. В модели сделано несколько общих для этих представлений предположений.

1) Ячейка может располагаться в одной или нескольких соседних строках и в одном или нескольких соседних столбцах (например, атрибуты COLSPAN и ROWSPAN в HTML) и имеет прямоугольную форму в пространстве строк и столбцов, как показано на рис. 1, *а*.

2) Внутри ячейки не могут располагаться другие ячейки или таблицы (это не допускается в Excel).

3) Содержимое ячейки может являться либо меткой (заголовком), либо вхождением (данными). Используемые здесь термины «вхождение» и «метка» соответствуют смыслу терминов «entry» и «label» соответственно из работы Wang X. [20].

4) Метки могут адресовать вхождения либо в строках — метки строк, либо в столбцах — метки столбцов.

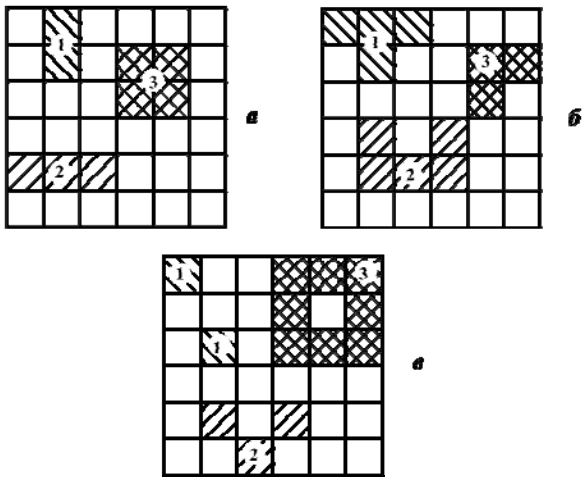

Рис. 1. Примеры объединения плиток сетки в ячейки таблицы, обозначенные как 1, 2 и 3: так ячейка может объединять несколько плиток в Excel, Word, HTML и LaTeX (*а*); так ячейка может визуально (для восприятия человеком) включать несколько плиток с помощью разграфки (*б*); скорее всего, так ячейки никто не представляет (*в*)

Очевидно, что сделанные предположения описывают широкий класс обрабатываемых таблиц. Пример сводной таблицы, полностью укладывающейся в данную модель, приводится на рис. 2.

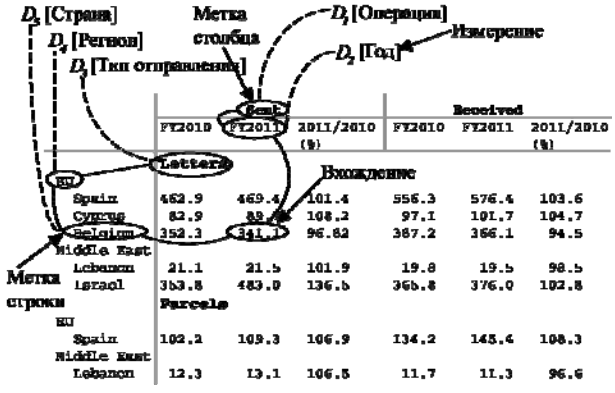

Рис. 2. Пример сводной таблицы

3.2 Модель таблицы

Модель включает два уровня: физической и логической структуры, которые в упрощенном виде можно описать следующим образом.

1) Уровень физической структуры $Tp=(Sr, Sc, C)$ состоит из: (1) пространства строк — Sr и столбцов — Sc; (2) набора ячеек — C, в котором каждая ячейка — $c=(p, c', S)$ включает: координаты в пространстве строк Sr и столбцов Sc — $p=(cl, rt, cr, rb)$, содержание — c', стилевая информация (цветовые схемы, шрифтовые метрики, выравнивание, стили оформлении границ и др.) — S.

2) Уровень логической структуры $Tl=(D, Lr, Lc, E)$ состоит из: (1) набора представленных в обрабатываемой таблице измерений — $D=\{Di\}$, каждое из которых содержит значения $Di=\{dj\}$; (2) дерева меток строк — Lr и (3) столбцов — Lc, отражающих связи между метками, не являющимися значениями измерений Di из набора D — $l=(l')$, где l' — содержание метки; (4) набора вхождений — E, в котором каждое вхождение — $e=(e', D', L')$ включает: содержание — e', набор связанных с ним значений измерений Di из набора D — D', набор связанных с ним меток из деревьев Lr и Lc — L'.

3.3 Структуры данных

Предлагаемая в работе модель таблицы реализована в виде ряда структур данных, основные из которых перечислены далее: CELL, ENTRY, LABEL, LABELNODE. Структура CELL предназначена для представления ячейки и прежде всего информации о её физической структуре, однако она также включает уровень логической структуры ячейки (т.е. она позволяет накапливать информацию о ее связях с другими ячейками, ее роли и типе данных). На практике это позволяет разрабатывать правила анализа табличной структуры в более лаконичной манере по сравнению со случаем, при котором используются дополнительные структуры данных для представления информации уровня логической структуры. Структуры ENTRY, LABEL, LABELNODE используются исключительно на уровне логической структуры. ENTRY служит для представления вхождения, а LABEL — метки. Структура LABELNODE является оболочкой для структуры LABEL и обеспечивает представление деревьев меток.

Все предложенные структуры данных и алгоритмы реализованы на платформе Java. Это обеспечивает возможность использовать их напрямую для представления фактов о таблицах в процессе логического вывода, выполняемого в системе исполнения правил с поддержкой спецификации JSR-94 (Java Rule Engine API).

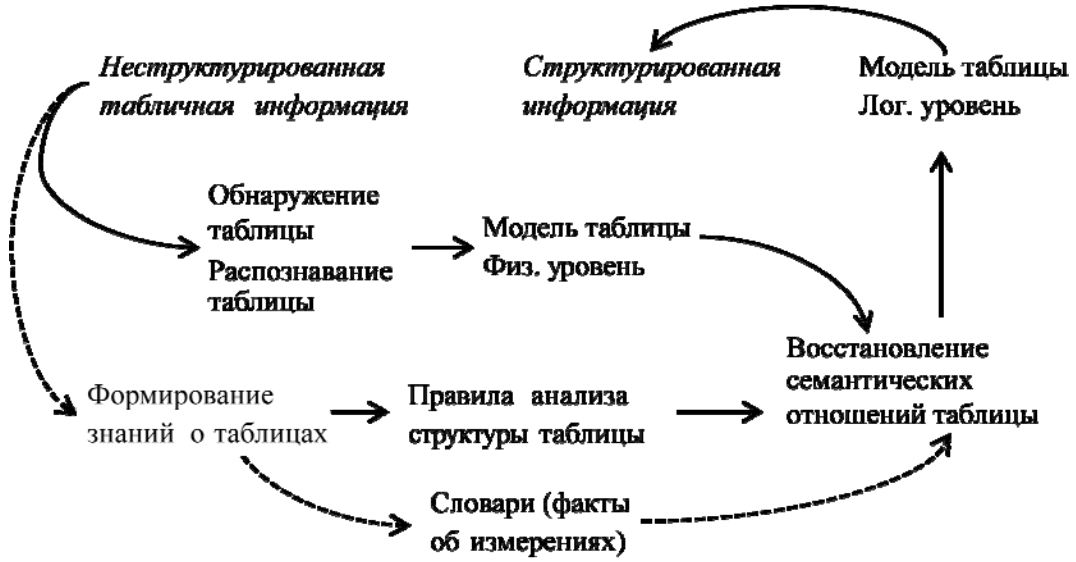

Рис. 3. Схема структурирования табличной информации

а

```
...
when
    $c : CCell( cl == 1, style.getFont().getColor() == "#ff0000")
then
    modify ( $c ) { setRole( Role.ROWLABEL ) }
```

б

```
...
when
    $c1 : CCell()
    $c2 : CCell( rt == $c1.rb + 1, ( $c1.cl <= cl && cr < $c1.cr ) ||
                ( $c1.cl < cl && cr <= $c1.cr ) )
then
    $c1.addConnectedCell( $c2 )
```

в

```
...
when
    $c : CCell( text matches "(?i).*(total)" )
then
    modify ( $c ) { setIgnored( true ) }
...
```

Рис. 4. Примеры правил анализа табличной структуры

4. Представление и исполнение правил анализа табличной структуры

Схема преобразования табличной информации от неструктурированной к структурированной форме показана на рис. 3. Предполагается, что этапы обнаружения и распознавания таблицы выполняются в сторонних системах. Например, для извлечения таблиц из PDF документов могут использоваться системы Tabula [18] или PDFGenie [15], для документов, напечатанных в файлы формата EMF, может использоваться технология, предложенная в работах [24]. Выходом таких систем являются таблицы в форматах Excel, HTML или XML, которые могут быть приведены к физическому уровню модели CELLS.

В процессе загрузки таблиц из полученных файлов Excel, HTML или XML в структуры данных, реализующих модель CELLS, табличная информация подвергается предобработки. Это включает опционально: удаление лишних пробельных и служебных символов из текстового содержания, исключение из таблицы пустых строк и столбцов и восстановление отсутствующих настроек стилей границ ячеек. Последнее необходимо, поскольку видимые и физические границы ячейки не всегда совпадают. Визуально они могут быть образованы границами соседних ячеек. Приведение стилей физических границ ячеек в соответствии с её

видимыми границами позволяет упростить правила анализа структуры таблицы.

Полученные в результате данные о таблице, которые формируют базу фактов для логического вывода. Кроме того, факты могут быть дополнены внешней информацией об измерениях.

Для обработки набора таблиц формируется база знаний, которая состоит из продукционных правил анализа табличной структуры. Они отображают доступную информацию: позиции (координаты), графическое форматирование и естественно-языковое содержание ячеек, в отсутствующие изначально отношения между метками, вхождениями и измерениями. Полученные в процессе вывода новые факты о семантических отношениях должны быть достаточными для каноникализации таблицы.

В качестве система исполнения таких правил может использоваться свободная системы Drools Expert [3], реализующая спецификацию JSR-94. При этом сами правила могут быть представлены на языке выражений MVEL [14].

На Рис. 4 приводится ряд простых примеров возможных правил анализа структуры на языке MVEL. Если ячейка $c находится в 1-ом столбце, а её текст выделен красным цветом, то она выполняет роль метки строки (рис. 4, *а*). Если ячейка $c1 расположена непосредственно над ячейкой $c2 и при этом полностью охватывает её по столбцам, то они связаны (рис. 2, *б*). Если ячейка $c содержит текст, удовлетворяющий регулярному выражению "(?i).*(total)", то её необходимо игнорировать при формировании выходных данных (рис. 2, *в*). Примеры правил, которые применялись при тестировании системы CELLS, можно найти по адресу http://cells.icc.ru/test.

В процессе логического вывода накапливается информация о логической структуре таблицы. Для этой информации выполняется постобработка, которая включает: приведение текстового содержания ячеек к эталонным написаниям, сопоставление меток с измерениями и формирование канонической формы таблицы.

Из восстановленной информации модели таблицы CELLS формируется таблица в канонической форме, которая включает следующие поля: DATA — данные (вхождения); ROW_LABEL — пути меток от листьев до корней из невырожденного дерева *Lr*; COL_LABEL — пути меток от листьев до корней из невырожденного дерева *Lc*; *D1,..., Dn* — поля значений измерений *Di* из набора *D*. Каждый кортеж в такой канонической форме представляет связь между вхождением, путями в деревьях меток и значениями восстановленных измерений. Дополнительно поле ROW_LABEL/COL_LABEL может быть разделено на несколько отдельных полей, каждое из которых будет соответствовать одному уровню вложенности в дереве меток строк/столбцов.

Пример канонической формы обработанной таблицы приводится на Рис. 5. Сформированная каноническая таблица может экспортироваться в реляционную базу данных с помощью стандартных средств интеграции данных известных систем управления базами данных (СУБД). Например, службы "SQL Server Integration Services" [17], позволяют импортировать данные из таблиц с простой "решеточной" структурой в форматах Excel, CSV в базы данных под управлением СУБД "SQL Server".

Данные	*Операция*	*Год*	*Тип отправления*	*Регион*	*Страна*
462.9	Sent	2010	Letters	EU	Spain
82.9	Sent	2010	Letters	EU	Cyprus
...	...	...	...	...	...
12.3	Sent	2010	Parcels	Middle East	Lebanon
469.4	Sent	2011	Letters	EU	Spain
89.7	Sent	2011	Letters	EU	Cyprus
341.1	Sent	2011	Letters	EU	Belgium
21.5	Sent	2011	Letters	Middle East	Lebanon
483.0	Sent	2011	Letters	Middle East	Israel
109.3	Sent	2011	Parcels	EU	Spain
13.1	Sent	2011	Parcels	Middle East	Lebanon
556.3	Received	2010	Letters	EU	Spain
...	...	...	...	...	...
11.3	Received	2011	Parcels	Middle East	Lebanon

Рис. 5. Каноническая форма таблицы из рис. 1: все метки сопоставлены измерениям, поэтому поля COL_LABEL и ROW_LABEL отсутствуют

3 Экспериментальные результаты

Экспериментальная оценка представленного подхода выполнена с помощью системы CELLS, в которой реализованы структуры данных, представляющие модель таблицы CELLS, и алгоритмы: 1) загрузки исходной табличной информации в формате Excel (тестовых данных со специальной разметкой); 2) структурирования табличной информации, восстановленной в процессе логического вывода; 3) экспорта результатов в формате Excel.

Для экспериментальной оценки сформирована коллекция тестовых данных, которая включает 97 таблиц в формате Excel, собранных из 7 различных источников. Коллекция доступна по адресу http://cells.icc.ru/test. Её краткое описание приводится в табл. 1. Для формирования коллекции исходная табличная информация была преобразована из формата PDF в Excel.

Источниками тестовых данных послужили слабоструктурированные документы в низкоуровневом формате файлов печати PDF — государственные и финансовые статистические отчеты с богатым табличным содержанием. Для формирования коллекции исходная табличная информация была преобразована из формата PDF в Excel. При этом, насколько это было возможно, в полученных тестовых таблицах было сохранено графическое форматирование, представленное в соответствующих им PDF источниках.

Таблица 1. Экспериментальные результаты

Код источника	Кол-во таблиц	Кол-во ячеек	Кол-во вхождений	Кол-во меток	Кол-во связей между метками*	Кол-во правил	Время исполнения правил (мс)
JAPAN_STAT [1]	15	1088	734	257	102	10	417
AEROFLOT [2]	13	2047	727	321	167	16	526
BOEING [3]	21	2156	964	470	196	14	663
CHINA_STAT [4]	18	7216	4180	862	551	12	964
CHEVRON [5]	7	812	268	141	89	12	283
USDA_NASS [6]	7	1553	1175	313	174	16	638
TOBACCO [7]	16	2844	2195	508	335	10	730

[1] Statistical Handbook of Japan 2007. Statistics Bureau of Japan. Chapter 5, 8.
[2] OJSC «Aeroflot – Russian Airlines» Consolidated Financial Statements For the Year Ended December 31, 2006. P. 4–10, 25–26.
[3] Boeing Co, Annual Report 2010. P. 50–55, 83–85.
[4] China statistical yearbook 2003. National Bureau of Statistics of China. P. 23–48, 555, 559, 571, 584, 590, 664, 708, 774, 765.
[5] Chevron Corp. News Release November 2, 2012. Chevron Corp. P. 1, 5–9.
[6] USDA NASS. 2003 Agricultural Statistics Annual. USDA (U.S. Department of Agriculture). National Agricultural Statistics Service. Chapter VI. P. 5–7, 12.
[7] Tobacco: World Markets and Trade 2005. USDA (U.S. Department of Agriculture). Foreign Agricultural Service.
* Исключая связи корней деревьев меток.

Тестовые данные имеют дополнительную разметку для определения местоположения таблицы внутри листа Excel (рис. 6), а также аккуратную декомпозицию на ячейки. Там, где это возможно, их физическая структура и разграфка совпадают. Это позволяет избежать этапов обнаружения и сегментации таблицы.

$START						
	Company name	Place of incorporation and operation	Activity	Percentage held as of December 31, 2006	Percentage held as of December 31, 2005	
	LLC "Airport Moscow"	Moscow region	Cargo handling	50,00%	50,00%	
	CJSC "Aerofirst"	Moscow region	Trading	33,30%	33,30%	
	CJSC "TZK Sheremetyevo"	Moscow region	Fuel trading company	31,00%	31,00%	
	CJSC "AeroMASH – AB"	Moscow region	Aviation security	45,00%	45,00%	
						$END

Рис. 6. Дополнительная разметка тестовой таблицы: маркеры «$START» и «$END» указывают соответственно верхний левый и нижний правый угол таблицы в пространстве строк и столбцов

На рис. 7 показаны некоторые таблицы из тестовой коллекции данных. Их структуры включают типичные для этой коллекции особенности. Так, таблица, рис. 7, *а*, содержит иерархии меток строк и столбцов. Тело таблицы, рис. 7, *б*, пересекают перерезы: «Price per 100 pounds» и «Price per bushel». В таблице, рис. X, *в*, столбцы с метками строк чередуются со столбцами с данными.

Полученные экспериментальные результаты приводятся в табл. 1. Логический вывод выполнялся в системе Drools Expert (5.4.0.Final). При этом использовался процессор Intel Core 2 Quad, 2,66 ГГц. Экспериментальные результаты показывают эффективность применения предлагаемого подхода для широкого класса таблиц.

4. Заключение

Предлагаемый подход базируется на предположении о том, что для одного или нескольких схожих источников можно разработать непротиворечивый набор правил анализа структуры содержащихся в них таблиц. Однако разработка достаточно универсальных баз знаний для многих разнородных источников имеет слишком высокую цену и не всегда возможна из-за противоречий, содержащихся в самих источниках. Поэтому данный подход предназначен в основном для задач управления данными, прежде всего для массовой интеграции табличной информации из наборов похожих источников.

Item	Total	National forest	Non-national forest		
			Municipal	Private	Others
Forest land area (1,000 ha)	25 121	7 838	2 796	14 440	46
Forest growing stock (1 mil. m3)	4 040	1 011	433	2 590	5
Planted forests					
Land area (1,000 ha)	10 361	2 411	1 232	6 705	12
Growing stock (1 mil. m3)	2 338	368	255	1 712	3
Natural forests					
Land area (1,000 ha)	13 349	4 770	1 426	7 126	27
Growing stock (1 mil. m3)	1 701	642	178	878	3

а

Kind of seed	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
	Price per 100 pounds									
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
Alfalfa, uncertified varieties	152.00	161.00	168.00	185.00	185.00	205.00	184.00	165.00	158.00	280.00
Alfalfa, certified varieties	269	266	274	277	282	288	287	277	278	157
Clover, ladino	324	321	320	318	307	308	298	285	285	280
Clover, red	148	148	134	172	184	194	178	143	132	130
Lespedeza, Korean	132	84,5	66	99	90	89	76,15	77,5	160	98
Sunflower	300	297	297	313	355	380	400	395	407	407
Cottonseed, all	62,7	63,5	68,2	73	74,9	79,3	82,4	128	154	213
Biotech [1]									217	271
Non-biotech									87	94
Grain sorghum, hybrid	74,5	82,1	78,7	84	92	96	97,6	93	93	96
	Price per bushel									
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
Corn, hybrid, all [2]	72,7	73,4	77,1	77,7	83,5	86,9	88,1	87,5	92,2	92
Biotech [1]									110	113
Non-biotec									85,3	85,8
Wheat (spring)	5,98	7,37	7,12	8,1	7,3	6,85	6,1	6,1	6,2	6,5
Wheat (winter)	7,73	7,9	7,8	8,5	10	8,25	7,35	7,05	7,2	7,7
Rice	15,4	22	15,1	17,5	19	19,5	19,1	17,25	15,7	14,9
Barley (spring)	5	5,18	5,37	6,49	6,13	6,04	5,8	5,8	5,8	5,8
Soybeans for seed, all	12,4	13,6	13,4	14,8	16,1	17,15	17	17,1	20,7	22,5
Biotech [1]									23,9	27
Non-biotec									17,9	15
Flaxseed	7,37	7,74	8	8,14	9,31	10	8,5	7,9	7,6	7,6

б

线路名称	Name	客运量（万人）Passenger Traffic (10 000 persons)	旅客周转量（百万人公里）Passenger-kilometers (million passenger-km)	线路名称	Name	货运量（万吨）Freight Traffic (10 000 tons)	货物周转量（百万吨公里）Freight Ton-kilometers (million ton-km)
京沪线	Beijing-Shanghai	5496	32975	京沈线	Beijing-Shenyang	3438	82790
新石线	Xinjiang-Rizhao			哈大线	Harbin-Dalian	3233	60717
沪杭	Shanghai-Hangzhou	654	6188	津沪线	Tianjin-Shanghai	5304	100909
浙赣线	Hangzhou-Ganzhou	3785	33028	沪杭线	Shanghai-Hangzhou	202	4939
鹰厦线	Yingtan-Xiamen	10869	88717	京广线	Beijing-Guangzhou	7187	131196
京九线	Beijing-Kowloon	814	1906	南北同蒲线	Datong-Taiyuan-Fenglingdu	11168	30412
京广线	Beijing-Guangzhou	491	1708	太焦柳线	Taiyuan-Jiaozuo-Liuzhou	8206	56729
石太线	Shijiazhuang-Taiyuan	1800	9364	京九线	Beijing-Kowloon	2644	61919
石德线	Shijiazhuang-Dezhou	1575	6452	兰新线	Lanzhou-Urumqi	3366	63348
焦柳线	Jiaozuo-Liuzhou	655	2512	滨洲线	Harbin-Manzhouli	3137	21181
京包线	Bingjing-Baotou	541	1288	滨绥线	Harbin-Suifenhe	1178	16384
包兰线	Baotou-Lanzhou	1245	3615	京包线	Bingjing-Baotou	5881	57077
北同蒲线	Taiyuan-Datong	4759	33838	石太线	Shijiazhuang-Taiyuan	3760	21301
南同蒲线	Fenglingdu-Taiyuan	74	1459	石德线	Shijiazhuang-Dezhou	379	11664
陇海线	Lianyungang-Lanzhou	1055	16149	浙赣线	Hangzhou-Ganzhou	2464	45035
宝中线	Baoji-Zhongwei	413	1865	陇海线	Lianyungang-Lanzhou	6357	100027

в

Рис. 7. Примеры тестовых таблиц

Подход положен в основу развиваемой авторами системы понимания таблиц в формате Excel. Полученные экспериментальные результаты показывают эффективность её применения для широкого класса таблиц, представленных в формате Excel. В то же время необходимо дальнейшее исследование возможностей для упрощения правил анализа структуры таблицы за счет развития структур данных представления табличной информации и дополнительных алгоритмов её предобработки и постобработки.

Работа выполнена при финансовой поддержке РФФИ грант № 14-07-00166 и Совета по грантам Президента РФ СП-3387.2013.5.

Литература

[1] Blumberg R., Atre S. The problem with unstructured data // DM Review, 2003. http://soquelgroup.com/Articles/dmreview_0203_problem.pdf

[2] Douglas S., Hurst M., Quinn D. Using Natural Language Processing for Identifying and Interpreting Tables in Plain Text // Proc. of the 4th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas. 1995. P. 535–546.

[3] Drools Expert (JBoss Community). http://www.jboss.org/drools/drools-expert.html

[4] e Silva A.C., Jorge A.M., Torgo L. Design of an end-to-end method to extract information from tables // Int. J. on Document Analysis and Recognition. 2006. Vol. 8, No. 2. P. 144–171.

[5] Embley D.W., Hurst M., Lopresti D., Nagy G. Table-processing paradigms: a research survey // Int. J. on Document Analysis and Recognition. 2006. Vol. 8, No. 2. P. 66–86.

[6] Embley D.W., Tao C., Liddle S.W. Automating the Extraction of Data from HTML Tables with Unknown Structure // Data & Knowledge Engineering. Elsevier. 2005. Vol. 54, No. 1. P. 3–28.

[7] Feldman R., Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data // Cambridge University Press. 2006. 422 p.

[8] Gatterbauer W., Bohunsky P., Herzog M., Krüpl B., Pollak B. Towards Domain-Independent Information Extraction from Web Tables // Proc. of the 16th Int. Conf. on World Wide Web. ACM New York, NY, US, 2007. P. 71–80.

[9] Hurst M. Layout and Language: Challenges for Table Understanding on the Web // In Proc. of the 1st Int. Workshop on Web Document Analysis. 2001. P. 27–30.

[10] Hurst M. The Interpretation of Tables in Texts. PhD thesis. School of Cognitive Science, Informatics, the University of Edinburgh. UK, 2000.

[11] Inmon W.H. Matching unstructured data and structured data // The data administration newsletter. 2006. http://www.tdan.com/view-articles/5009.

[12] Inmon W.H., Nesavich A. "Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence", 1st edition, Prentice Hall PTR, 2007.

[13] Lopresti D., Nagy G. A tabular survey of automated table processing // Lecture Notes in Computer Science. 2000. Vol. 1941. P. 93–120.

[14] MVEL. http://mvel.codehaus.org

[15] PDFGenie, http://www.pdftron.com/pdfgenie

[16] Shilakes C.C., Tylman J. Enterprise Information Portals // Merrill Lynch. 1998.

[17] SQL Server Integration Services, http://msdn.microsoft.com/ru-ru/library/ms141026.aspx

[18] Tabula, http://tabula.nerdpower.org

[19] Tijerino Y., Embley D., Lonsdale D., Nagy G. Towards ontology generation from tables // World Wide Web: Internet and Web Information Systems. 2005. Vol. 8, No. 3. P. 261–285.

[20] Wang X. Tabular Abstraction, Editing, and Formatting. PhD thesis. Waterloo, Ontario, Canada. 1996.

[21] WordNet, http://wordnet.princeton.edu

[22] Zanibbi R., Blostein D., Cordy J.R. A survey of table recognition: Models, observations, transformations, and inferences // Int. J. on Document Analysis and Recognition. 2004. Vol. 7, No. 1. P. 1–16.

[23] Кудинов П.Ю. Адаптивные методы извлечения информации из статистических таблиц, представленных в текстовом виде : дис. … канд. техн. наук. М., 2011. С. 105.

[24] Шигаров А.О. Технология извлечения табличной информации из электронных документов разных форматов : дис. … канд. техн. наук. Иркутск, 2009. С. 143.

Automated Table Understanding Using a Rule Engine

Alexey O. Shigarov

The paper discusses issues on automation of the table understanding (i.e. recovering relationships of table elements). We propose an approach to table understanding based on the use of a rule engine. A table model oriented on the logical inference and algorithms for processing tabular information are also considered in the paper. The CELLS system for structuring tabular information presented in Excel spreadsheet format has been developed using the proposed approach, model and algorithms. The performance evaluation of the system shows that the approach can be applied to a wide range of tables.