Извлечение таблиц из неформатированного текста

Хмельнов А. Е., Шигаров А. О.
hmelnov@irk.ru, shigarov@icc.ru
Иркутск, Институт динамики систем и теории управления СО РАН

В данной работе представлены основные концепции разработанного нами эвристического метода извлечения таблиц из неформатированного текста. Метод использует особенности структуры статистических таблиц, публикуемых Росстатом. Также эти особенности в полной мере относятся к статистическим таблицам, представленным в государственных статистических отчетах США (www.fedstats.gov), Евросоюза (Eurostat yearbook 2006-07) и Японии (Statistical Handbook of Japan 2006). Метод может быть применен к подобным таблицам, представленным, как неформатированный текст.

Извлечение таблиц из документов является одной из задач, решаемых в системах анализа и обработки документов. Обзоры работ по данной проблеме [1, 2, 3], появившиеся за несколько последних лет, показывают растущий интерес к данной проблематике. В литературе выделяются следующие основные стадии обработки, которые могут быть выполнены при извлечении таблиц: обнаружение таблиц в документах, сегментация таблиц на отдельные клетки, функциональный анализ — определение роли клеток, структурный анализ — определение зависимостей между клетками, и интерпретация — преобразование табличной информации к требуемому виду. В работах [4, 5, 6] предложены различные подходы к обнаружению таблиц в неформатированном тексте. В работе [7] предложен метод извлечения таблиц из неформатированных текстов, в котором реализованы все вышеупомянутые стадии обработки, но при этом используются слишком сильные предположения о структуре обрабатываемых таблиц.

Как правило, методы извлечения таблиц из документов ориентируются на определенные среды и форматы представления документов, а также на определенную структуру таблиц, которая обычно определяется стандартами и соглашениями, принятыми в той предметной области, где используются эти таблицы.

В данной работе рассматривается метод, учитывающий особенности статистических таблиц, и позволяющий выполнить все стадии их обработки, результатом применения которого является извлечение информации из текстовых таблиц в реляционную БД. На Рис. 1 показан пример статистической таблицы. Рассматриваемые таблицы состоят из шапки и тела, кроме того, они могут иметь боковик и перерезы. Тело таблицы содержит только числовые данные. Заголовки столбцов обычно выделяются линейками, составленными из символов псевдографики или

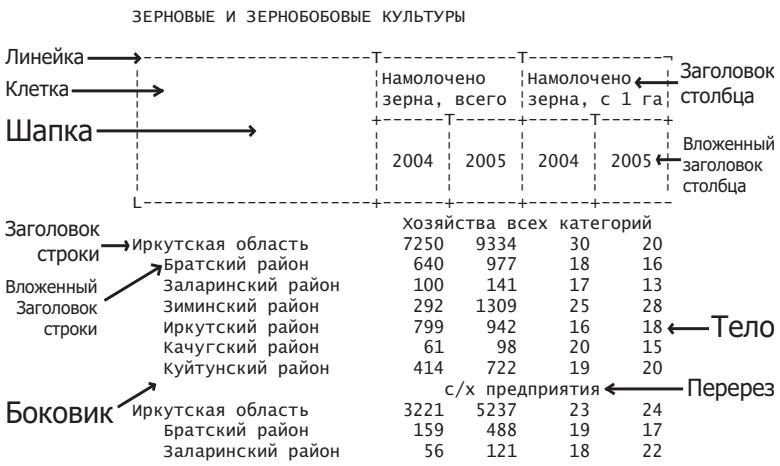

Рис. 1. Пример статистической таблицы

подобных им символов набора ASCII. Пересекаясь, линейки образуют клетки, которые ограничивают отдельные заголовки столбцов. Один или несколько заголовков могут быть вложенными в другой заголовок; в этом случае, клетки, ограничивающие их, лежат сразу под клеткой, ограничивающей заголовок, в который они вложены. Заголовки строк также имеют вложенность, которая определяется отступом от левого края таблицы.

Особенности компоновки заголовков столбцов позволяют представить шапку в виде дерева, узлами которого являются заголовки столбцов, а ребрами — пары заголовков (h_a, h_b), где h_a — заголовок, вложенный в h_b. Корнем этого дерева является пустой элемент, заголовки верхнего уровня являются его подузлами. Подобным образом, о боковике также можно думать, как о дереве, в котором заголовки строк являются узлами, а пары заголовков строк, в которых один вложен в другой — ребрами. Перерезы также удобнее рассматривать как дерево, хотя они не имеют вложенности.

В результате для представления рассматриваемых таблиц может быть предложена следующая модель. Пусть H^t, S^t и C^t — деревья, представляющие соответственно шапку, боковик и перерезы, а H, S и C — множества узлов, соответствующие этим деревьям. Пусть $V\colon V \subset \mathbb{R}$ — множество всех значений из тела таблицы. Пусть $L \subseteq H \times S \times C$ — подмножество таких элементов из $H \times S \times C$, для которых определено значение $v \in V$. Тогда множество $T = \{H^t, S^t, C^t, L \to V\}$ составляет модель таблицы.

Прежде всего, решается задача обнаружения таблиц в тексте. При этом используется ряд предположений о наличии в шапке таблицы символов-разделителей. После обнаружения и сегментации заголовка таблицы выполняется сегментация строк текста, составляющих боковик и тело, а также выделяются перерезы. Таким образом, приходим к описанной модели таблицы. Используемые предположения выполняются для абсолютного большинства рассматриваемых таблиц, в случае их нарушения результаты сегментации могут редактироваться пользователем.

Следующим шагом является классификация и нормализация обнаруженных узлов в деревьях заголовков. Для этих целей используются правила, задаваемые при помощи регулярных выражений. Могут выделяться значения, относящиеся к различным измерениям (временной интервал, территория), а также игнорируемые узлы. При обнаружении в заголовке значения, оно сопоставляется всем клеткам таблицы, подчинённым рассматриваемому узлу, а сам узел исключается из дерева (с сохранением подчинённых узлов).

На основе предложенного метода разработано приложение для перевода текстовых таблиц в реляционную БД, с использованием которого в сжатые сроки было обработано около 2800 таблиц, содержащих более 21000 показателей и более 300000 значений.

Литература

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

[2] *Zanibbi R., Blostein D., Cordy J. R.* A survey of table recognition: Models, observations, transformations, and inferences // International Journal on Document Analysis and Recognition. — 2004. — Vol. 7, No. 1. — pp. 1–16.

[3] *Embley D. W., Hurst M., Lopresti D., Nagy G.* Table-processing paradigms: a research survey // International Journal on Document Analysis and Recognition. — 2006. — Vol. 8, No. 2. — pp. 66–86.

[4] *Tupaj S., Shi Z., Chang C. H., Alam H.* Extracting Tabular Information From Text Files. — 1996. — `citeseer.nj.nec.com`.

[5] *Hu J., Kashi R., Lopresti D., Wilfong G.* Medium-Independent Table Detection // Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging), San Jose, 2000. — pp. 291–302.

[6] *Pinto D., McCallum A., Wei X., Croft B.* Table Extraction Using, Conditional Random Fields // 26th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval, 2003.

[7] *Douglas S., Hurst M., Quinn H.* Using Natural Language Processing for Identifying and Interpreting Tables in Plain Text // 4th annual symposium on document analysis and information retrieval, Las Vegas, 1995. — pp. 535–546.