Логический анализ компоновки таблиц из слабоструктурированных источников с использованием экспертных знаний

Шигаров Алексей Олегович
Институт динамики систем и теории управления СО РАН (Иркутск), Россия
e-mail: shigarov@icc.ru

Большое количество табличной информации представлено в неструктурированном (ASCII-текст, файлы печати) и слабоструктурированном (документы Word, HTML страницы, листы Excel) виде. Такая информация напрямую не предназначена для высокоуровневой машинной обработки, например, для преобразования к базам данных. Для того чтобы анализировать и обрабатывать большие объемы такой информации прежде всего необходимо представить ее в структурированном виде. Для такого преобразования требуется восстановление отсутствующей изначально в таблице семантической информации (структурных метаданных). Даже для достаточно высокоуровневого слабоструктурированного представления таблиц (например, в виде HTML разметки, объектов табличного процессора Excel или текстового процессора Word) необходимо восстанавливать отсутствующую информацию о разделении ячеек на атрибуты (заголовки) и данные, связях между ячейками, а также об используемых типах данных содержания ячеек (о лексических значениях, принадлежности измерениям). В литературе по системам анализа и распознавания документов эти задачи принято называть логическим анализом компоновки (*logical layout analysis*) [1] таблицы.

В данной работе предлагается проект системы трансформации таблицы от слабоструктурированного представления, содержащего информацию только о декомпозиции таблицы на ячейки, к отношению реляционной модели данных. Предлагаемая система обеспечивает анализ логической компоновки таблицы, в частности, разделение содержания ячеек на атрибуты и данные, восстановление отношений между атрибутами и значениями данных и отношения подчиненности между атрибутами, восстановление измерений (доменов). Рассматриваемый анализ ориентирован на таблицы со сложной компоновкой, являющиеся результатом использования систем генерации отчетов, сводных таблиц (pivot table) в табличных процессорах или OLAP (Online Analytical Processing). Их компоновка, как правило, отличается от простой “решеточной” компоновки, используемой по умолчанию в листах Excel и документах Word, наличием иерархий заголовков и нестрогим расположением заголовков и данных. Часто таблицы в рамках одной предметной области, корпоративной среды, организации, издания, отчета и т.д. строятся по определенным требованиям оформления. В предлагаемой системе такие правила представляются в продукционном виде с помощью системы Drools Expert [2]. При анализе логической компоновки заданные в базе знаний решающие правила отображают позиции элементов таблицы (расположение относительно друг друга и таблицы), их графическое форматирование и естественно-языковое содержание в отсутствующую изначально информацию — семантические отношения между элементами таблицы, их роли (атрибуты или данные) и типы данных (измерения).

Работа выполнена при поддержке РФФИ, гранты №№ 12-07-31051, 11-07-92204.

1. Machine Learning in Document Analysis and Recognition // Series: Studies in Computational Intelligence, Vol. 90. Marinai S., Fujisawa H. (Eds.). Springer-Verlag Berlin Heidelberg. 2008. 434 p.

2. Drools Expert, http://www.jboss.org/drools/drools-expert.html