МЕТОДЫ РАСПОЗНАВАНИЯ СТРУКТУРЫ ВЕБ-ТАБЛИЦ

Е. В. Горбач

Белорусский государственный университет, Минск, Беларусь
E-mail: katerinagorbac@gmail.com

В работе рассматриваются вопросы обнаружения семантической структуры таблиц. В связи с этим решается задача классификации типа таблицы, как наиболее популярный подход в распознавании структуры. Проведен сравнительный анализ современных методов, используемых при решении данной задачи. Разработан вариант улучшеной нейросетевой архитектуры, который показал увеличение точности на эталонной коллекции таблиц.

Ключевые слова: *веб-таблицы, автоматическая обработка документов, распознавание структуры веб-таблицы*

Таблицы являются распространенным инструментом отображения информации, так как человек может быстро и наглядно их интерпретировать. Автоматическая обработка таблиц также может быть полезна при сборе и хранении данных, которые традиционно представлены в полу структурированном виде: квитанции, чеки. Много данных такого формата не имеют однотипную структуру и нуждаются в предобработке для последующего анализа и хранения.

Сложности, которые возникают при обработке таблиц, связаны с тем, что таблицы содержат данные на естественном языке, а также имеют структуру, которую необходимо уметь определять. На данный момент нет решения «из коробки», которое смогло бы определить структуру любой семантически правильной таблицы. При этом для человека эта задача ясна и интуитивно понятна.

ПОНЯТИЕ СТРУКТУРЫ И ТИПА ТАБЛИЦЫ

Таблица представляет собой сетку ячеек, расположенных в строках и столбцах. Выделяют заголовок – список меток, определяющих содержание каждой строки/столбца таблицы. Заголовки обычно располагаются в первой строке/столбце таблицы и могут быть многоуровневыми.

Таблица хранит информацию о множестве объектов с определенными свойствами (атрибутами). Каждая запись может быть преобразована в пару «атрибут–значение». Названия свойств указываются в заголовках столбцов, а значения в ячейках. Свойства могут объединяться в иерархии и быть отображены через различные варианты визуализации.

В литературе предложен ряд схем классификаций типов таблиц. Мы будем использовать классификацию, предложенную в статье [1]. Эта

классификация разделяет таблицы по ориентации: вертикальная, горизонтальная и матричная.

Решая задачу определения класса таблицы (ее ориентации) для простых таблиц без иерархических заголовков, мы можем решить задачу определения семантической структуры таблицы и извлечь из таблицы пары «атрибут-значение».

В данной работе рассматриваются именно простые таблицы без иерархических заголовков, так как в открытых источниках отсутствуют размеченные коллекции таблиц с иерархическими заголовками.

КОЛЛЕКЦИЯ ДАННЫХ

Для оценки качества алгоритмов были использованы две коллекции таблиц (таблица 1):

1. Коллекция, далее именуемая как коллекция ScienceTable, представленная в работе [2] для оценки модели DeepTable, основаная на данных из научных статей. Разметка данного корпуса была получена эвристически на основании признаков html разметки.

2. Авторская коллекция таблиц, полученная из веб-страниц по налогообложению, далее именуемая коллекция TaxTable. Коллекция размечена вручную.

Таблица 1

Количество таблиц по классам

Класс	Коллекция ScienceTable	Коллекция TaxTable
Горизонтальная ориентация	1835	40
Вертикальная ориентация	1835	40
Матричная ориентация	1835	20

МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ ТИПА ТАБЛИЦЫ

Классический подход решения задачи классификации типа таблицы – генерация признаков, основанных на длине текста в ячейке и содержания ячейки [3,4] и использование классификатора поверх этих признаков. В данной работе предлагается к рассмотрению подход, кодирующий распределение длины ячейки по строкам и столбцам, при помощи двойного применения комбинаций функций среднего, медианы, дисперсии. В отличии от известных подходов, предложенный алгоритм способен учитывать информации из всей таблицы. Обычно используется среднее и дисперсия длины ячейки в первом столбце/строке. Ниже в таблице 2 перечислены используемые признаки.

Таблица 2

Признаки для классификации

Признак	Описание	Подробности
% is num	Процент ячеек, состоящих только из числа	Применяется к 1 и 2 строке/столбцу
% contains num	Процент ячеек, содержащих число	Применяется к 1 и 2 строке/столбцу
% is unique	Процент уникальных ячеек	Применяется к 1 и 2 строке/столбцу
Std len	Стандартное отклонение длины строки в ячейке	Применяется ко всем строкам/столбцам; Для получения итогового признака к полученному вектору применяется std/mean/median
Mean len	Среднее длины строки в ячейки	См. выше
Median len	Медиана длины строки в ячейке	См. выше

В последнее время методы глубокого обучения также использовались для классификации типа таблицы [2]. Поскольку авторы [2] используют открытую коллекцию таблиц и их эксперименты воспроизводимы, в качестве бенчмарка в данной работе была использована архитектура DeepTable [2]. Она включает следующие операции: получение векторного представление ячейки (токены ячейки кодируются предобученными эмбедингами и затем подаются на вход LSTM); получение векторного представления отдельно по колонкам и столбцам; конкатенацию данных представлений; классификацию типа таблицы.

Использование в качестве классификатора дерева решений с признаками, перечисленными в таблице 1, дает прирост в точности на 5% в сравнении с результатами полученными авторами [2] на коллекции ScienceTable. Поэтому в качестве улучшения архитектуры DeepTable было использовано следующее векторное представление содержания ячейки: длина содержимого ячейки, является ли ячейка числом, содержит ли ячейка число, является ли ячейка уникальной. Данная архитектура показала прирост в точности на 6% в сравнении с результатами DeepTable (таблица 3). Преимуществом данной модификации в сравнении с исходной архитектурой DeepTable является отсутствие необходимости использовать тяжеловесные предобученные эмбединги и значительное уменьшение числа настраиваемых параметров модели.

Таблица 3

Сравнение точности на отложенной выборке в рассмотренных методах

Коллекция	Случайный выбор	Решающее дерево*	DeepTable**	DeepTable с визуальными признаками*	DeepTable с визуальными признакми дообученный на TaxTable*
DeepTable	34.1%	78.7%	73.41%	**79.18%**	-
TaxTable	20%	**72%**	-	50%	50%

* - решение, предложенное в данной работе,
**- результаты, приведенные в работе [2]

Эксперименты на коллекции ScienceTable показали, что пространственные признаки вносят больший вклад в понимание класса таблицы в сравнении с семантическими признаками.

Также по полученным результатам видно, что использование нейронной сети, обученной на коллекции ScienceTable с использованием визуальных признаков, не позволяет без дообучения получать хорошие результаты на коллекции TaxTable.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Lehmberg O., Ritze D., Meusel R., Bizer C. 2016. A Large Public Corpus of Web Tables containing Time and Context Metadata. // International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. P. 75–76. DOI:10.1145/2872518.2889386.
2. Habibi M., Starlinger J., Leser U. DeepTable: a permutation invariant neural network for table orientation classification. // Data Min Knowl Disc 34. P. 1963–1983. DOI: 10.1007/s10618-020-00711-x.
3. Wang Y., J. Hu. A machine learning based approach for table detection on the web. // of the 11th International Conference on World Wide Web. 2002. P. 242-250.
4. Crestan E., P. Pantel. Web-scale table census and classification. // In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). Association for Computing Machinery, New York, USA. 2011. P. 545–554. DOI:https://doi.org/10.1145/1935826.19359042011
5. Zhang S., Balog K. Web Table Extraction, Retrieval and Augmentation: A Survey // ACM Transactions on Intelligent Systems and Technology. 2020. V. 11. N. 2.