УДК 004.91

Разработка метода унификации форматов изоморфных табличных данных

Сапельникова К.С. (ИТМО), **Першинов А.В.** (ИТМО), Басилаев Д. И.
Научный руководитель – кандидат технических наук, доцент Насонов Д.А.
(ИТМО), **инженер, аспирант Ковальчук М.А.**
(ИТМО)

Введение. На данный момент в мире наблюдается стремительный рост объема информации, представленной в различных формах. Одной из таких форм являются документы, содержащие ценную для отрасли информацию, но анализ таких документов затруднен по нескольким причинам. Во-первых, на ввод и очистку данных уходит слишком много времени и ресурсов. Кроме того, данные содержат несоответствия и ошибки. Таким образом, из-за низкого качества входных данных снижается точность анализа.

Следствием этого является задача обработки документов, первым шагом которой является обработка документов, содержащих специализированную терминологию. За этим следует этап извлечения структуры и ключевых объектов документа. Наконец, фрагментация содержимого документа, которая рассматривается в данной работе.

Основная часть. Анализ существующей работы показал, что существуют три большие группы задач, связанных с таблицами в документах. Первая группа включает обнаружение таблиц в документах, которые не поддерживают табличные форматы, такие как pdf-файлы или изображения. Для решения этой задачи были разработаны такие модели, как EDD[1] и CascadeTabNet[2]. Ко второй группе относятся работы, связанные с изучением структуры таблиц. В частности, TableSense[3] справляется с поиском таблицы на листе, а TUTA[4] извлекает и структурирует заголовки сложных таблиц. Однако ни одна из этих моделей не имеет дело с содержимым таблицы. Третья группа задач связана с анализом содержимого таблиц. В частности, TAPAS[5] — это модель для ответов на вопросы о содержании таблиц на естественном языке. Однако эти модели работают с анализом отдельных таблиц.

Для данной работы был собран набор данных, состоящий из таблиц, содержащих аналогичные данные в изоморфном формате. Кроме того, для приближения к реальным условиям были выбраны таблицы большого объема, которые содержали такие сложности, как неправильный формат или тип данных, сложный заголовок, подзаголовки, объединенные ячейки и пропущенные значения.

В качестве входных данных для разработанного алгоритма подается таблица с отмеченными границами. Каждая ячейка таблицы векторизуется и описывается набором ее признаков. Кроме того, в процессе векторизации определяется тип строки исходной таблицы. Это необходимо для определения подзаголовков. После этого результирующий набор данных делится на части в зависимости от типов данных ячеек. Далее часть набора данных кластеризуется, а затем происходит автоматическая маркировка кластера с использованием частичной маркировки, сделанной пользователем. После чего планируется восстановить структуру таблицы в унифицированном формате.

Выводы. В ходе данной работы был собран набор данных, определен набор наиболее важных признаков, реализован анализ типов строк для обнаружения подзаголовков в теле таблицы, а также разработан алгоритм унификации форматов изоморфных табличных данных.

Список использованных источников:

1) X. Zhong, E. ShafieiBavani, and A. Jimeno Yepes, “Image-based table recognition: Data, model, and evaluation,” in Computer Vision – ECCV 2020 (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 564–580, Springer International Publishing, 2020.
2) Jiayi Yuan, Hongye Li, Meng Wang, Ruyang Liu, Chuanyou Li and Beilun Wang, "An OpenCV-based Framework for Table Information Extraction", 2020 IEEE International Conference on Knowledge Graph (ICKG), vol. 621, pp. 628, 2020.
3) H. Dong, S. Liu, S. Han, Z. Fu, and D. Zhang, “Tablesense: Spreadsheet table detection with convolutional neural networks,” in AAAI, 2019.
4) Wang, Z., Dong, H., Jia, R., Li, J., Fu, Z., Han, S., & Zhang, D. (2021). Tuta: Tree-based transformers for generally structured table pre-training. In Proceedings of 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 1780–1790).
5) Herzig, J.; Nowak, P. K.; M¨uller, T.; Piccinno, F.; and Eisenschlos, J. M. 2020. TAPAS: Weakly Supervised Table Parsing via Pre-training. arXiv preprint arXiv:2004.02349 .