ОБЗОР АЛГОРИТМОВ ВОССТАНОВЛЕНИЯ ПОРЯДКА ЧТЕНИЯ

Карпенков М. А., Копылов Д. Е.

Ключевые слова: восстановление порядка чтения, структурирование текста, логическая структура документа, анализ документов.

Восстановление порядка чтения – ключевая задача анализа документов, востребованная при обработке научных статей, газет, технической документации и других структурированных текстов. Рассмотрены четыре алгоритма решения задачи, метрики оценки качества, а также наборы данных для будущего сравнительного анализа.

Постановка задачи: Для множества текстовых блоков B = $\{b_i\}$, где $b_i = [x_1, y_1, x_2, y_2]$ – координаты ограничивающей рамки, требуется определить порядковый номер чтения каждого элемента. Допускается использование меток $l \in L$ (например, title, img, body_of_text).

Было рассмотрено 4 алгоритма:

1. Геометрические правила [2]. Алгоритм сортирует блоки по вертикальному и горизонтальному расположению. В силу простоты алгоритм эффективен только для документов с сильной многоколоночной структурой

2. Алгоритм в работе [1] основан на отношениях Аллена, которые включают 13 интервальных отношений на осях X и Y. Эти отношения используются для определения булевой функции $f(a, b)$ которая определяет, читается ли блок a раньше блока b.

3. Алгоритм в работе [4] основан на теории аргументации. В качестве аргументов выступают утверждения о порядке чтения двух блоков. Сам алгоритм работает с конфликтами (атаками, противоречиями) аргументов, выбирая из всех возможных аргументов неконфликтующие и охватывающие все блоки.

4. Алгоритм [3] разделяет документ на части с помощью горизонтальных и вертикальных прямых, которые не пересекаются с блоками. В результате создается дерево: корень – множество всех блоков, ветви – непересекающиеся подмножества после каждого сечения, листья – отдельные блоки. По этому дереву и координатам блоков восстанавливается порядок чтения.

Для оценки качества алгоритмов могут использоваться следующие метрики. Метрика BLEU измеряет количество совпадений между предсказанным порядком и правильным. Метрика ARD является суммой

отклонений номеров блоков в предсказанном порядке и их номеров в правильном.

Для анализа предложены следующие наборы данных: ReadingBank, содержащий 500 тысяч изображений документов с широким спектром типов документов, а также соответствующую информацию о порядке чтения, ROOR, содержащий около 400 страниц размеченных документов, MTDB, содержащий 171 размеченную страницу: технические журналы, газеты, UW-II, содержащий более 2000 страниц на английском: статьи, журналы, академические тексты.

Список литературы

1. Aiello M., Smeulders A. M. W. Bidimensional Relations for Reading Order Detection // EPRINTS-BOOK-TITLE / University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science. 2003. 4 pp.
2. Breuel T. High Performance Document Layout Analysis // University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science. 2003. Vol. 5. P. 209-218.
3. Gu Z., Meng C., Wang K., Lan J., Wang W., Gu M., Zhang L. XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding. 2022. https://arxiv.org/abs/2203.06947
4. Ferilli S., Pazienza A. An Abstract Argumentation-based Strategy for Reading Order Detection. 2014. 12 pp.