1C-7

Detection of Paragraph Boundaries in Complex Page Layouts for Electronic Documents

Chu Yimin † Jun Adachi †† Atsuhiro Takasu ††
†University of Tokyo ††National Institute of Informatics

1 Introduction

PDF(Portable Document Format) is a popular file format to store, present and destribute the electronic documents. There have been a few researches on the topic of retrieving the content from PDF files, such as [1]. PDF format itself has the congenital disadvantage of being difficult to retieve the content from the file for data mining and analysis tasks.

The text content in PDF file would lose most semantic information during conversion. In order to repurpose PDF documents for flexible presentation and content retrieval, it is usually needed to extract the content from the file and to segment the content, for the sake of further semantic restoration and analysis. Finding the correct boundary of the text content blocks would greatly affect the precision in later processing stages.

Recenty J. Fan proposed a new method for text segmentation in PDF format[2], which is reported as both flexible and robust. His method took font size and line space into considertion as the key visual features to determine the boundaries of text blocks. After examining the method and the experiment result, we found that another important visual layout feature, the indentation, was not counted at all. As a result, Fan's experiment result showed the tendency of grouping multiple paragraphs together, when there is no significant bigger spacing between them. Therefore, we want to extend this algorithm for automatic content extraction and segmentation from PDF files.

In this paper, we extended J. Fan's method of text segmentation. We added an indentation detection step along the original algorithm, in order to detect the indentations which are the clues of new paragraphs, while not importing recognition errors on other layout patterns. In section 2 we will describe our method in details. Section 3 the comparision of the segmentation results will be presented. Section 4 concludes the paper.

Detection of Paragraph Boundaries in Complex Page Layouts for Electronic Documents
† Chu Yimin(chuym1983@nii.ac.jp)
†† Jun Adachi(adachi@nii.ac.jp)
†† Atsuhiro Takasu(takasu@nii.ac.jp)
2-1-2, Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan

2 Proposed Method

Compared to the font size and line spacing, indentation itself is not sufficient to be the criterion to determine the boundary of paragraphs. Apart from the indentation at the start of a paragraph, there are also indentations before a list item. An indented edge also happens when the text block is aligned to the other side as Figure 1 shows. We process the blocks those are extracted by J. Fan's method in two steps: first we detect the alignment of the blocks, then we run a detection of a zig-zag pattern on the aligned side of the block.

2.1 Alignment Detection

In J. Fan's original method, the outline of an extracted block is a polygon. Here we use a bounding box for every block. A bounding box is a box on the page which contains just all the text content in this block. The blocks are extracted by J. Fan's algorithm. To simplify the discussion, we assume that the text direction on the page is horizonal.

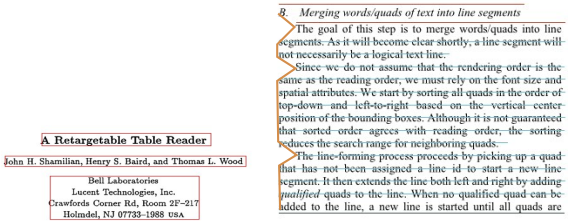
We indicate the alignment by calculating the sum of squres of the difference between the edge of each line and the edge of the block, as well as the center point of the line and that of the block. On an aligned edge, the sum of squre of the indentations will be significantly smaller than that on an unaligned edge.

2.2 Indentation Detection

When we have detected the alignments for the extracted text blocks, weshould run the indentation detection only on the aligned side of the block, which reduces the complexity and avoids the ambiguity of the meaning of the indentation.

Our method is described as below.

- For each line in the block
  - record the difference between current line



(a) A center aligned example

(b) The indented edge as the clue of paragraph segmentation

Fig 1: Indentation in different situations

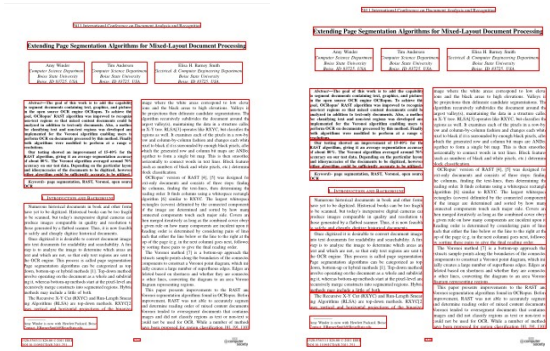
and last line onthe bounding box edges

$$\Delta E_i = L_i.left - L_{i-1}.left$$

- if the difference is positive, there is a new indentation appears. Before having premature decision, push the difference and the line number into a stack. Initial a step counter.
- else if the difference is close to zero, current indentation is being kept. Count the step counter.
- the last situation is that the difference is negative. It indicates that the indentation is being replaced. We keep poping the stack, until the previous indentation are covered by this negative difference. If the step counter is 1, and the stack is empty after pop, it matches the feature we observed from the indentations of a new paragraph. We reset the step counter and divide the block.
- in other cases, either the indentation is caused by list or novel layout settings. We would not divide the block.

3 Evaluation

As for evaluation, we randomly chose PDF files of proceeding papers in recent ICDAR and DAS. We ran both Fan's original method and our proposed method on the pages for comparison. Then we count the paragraphs our method extracted from the pages, compared to the numbers of paragraphs on those pages. An example of result is shown as



(a) (b)

Fig 2: Result of page segmentation

Figure 2. (a) shows the segmentaion after Fan's algorithm, while (b) presents the result after proposed algorithm. Our method trims and divides the text blocks into the paragraphs, while not compromising the precision Fan's algorithm achieved.

4 Conclusion

In this paper, we proposed the method of combining an existed text segmentation algorithm with the line indentation detection. From the result of experiment on paper pages, the segmentation result is robust and satisfactory. Based on the result of paragraph segmentation, we can pursue further research of extracting the sementic from the electronic documents.

Acknowledgment

Chu Yimin is funded by China Scholarship Council.

References

[1] Hui Chao and Jan Fan, "Layout and content extraction for PDF documents", S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 213-224, 2004

[2] Jian Fan, "Text Segmentation of Consumer Magazines in PDF Format", Proc. ICDAR, pp. 794-798, 2011