World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:11, 2022

An Observation Approach of Reading Order for Single Column and Two Column Layout Template

Authors : In-Tsang Lin, Chiching Wei

Abstract : Reading order is an important task in many digitization scenarios involving the preservation of the logical structure of a document. From the paper survey, it finds that the state-of-the-art algorithm could not fulfill to get the accurate reading order in the portable document format (PDF) files with rich formats, diverse layout arrangement. In recent years, most of the studies on the analysis of reading order have targeted the specific problem of associating layout components with logical labels, while less attention has been paid to the problem of extracting relationships the problem of detecting the reading order relationship between logical components, such as cross-references. Over 3 years of development, the company Foxit has demonstrated the layout recognition (LR) engine in revision 20601 to eager for the accuracy of the reading order. The bounding box of each paragraph can be obtained correctly by the Foxit LR engine, but the result of reading-order is not always correct for single-column, and two-column layout format due to the table issue, formula issue, and multiple mini separated bounding box and footer issue. Thus, the algorithm is developed to improve the accuracy of the reading order based on the Foxit LR structure. In this paper, a creative observation method (Here called the MESH method) is provided here to open a new chance in the research of the reading-order field. Here two important parameters are introduced, one parameter is the number of the bounding box on the right side of the present bounding box (NRight), and another parameter is the number of the bounding box under the present bounding box (Nunder). And the normalized x-value (x/the whole width), the normalized y-value (y/the whole height) of each bounding box, the x-, and y- position of each bounding box were also put into consideration. Initial experimental results of single column layout format demonstrate a 19.33% absolute improvement in accuracy of the reading-order over 7 PDF files (total 150 pages) using our proposed method based on the LR structure over the baseline method using the LR structure in 20601 revision, which its accuracy of the reading-order is 72%. And for two-column layout format, the preliminary results demonstrate a 44.44% absolute improvement in accuracy of the reading-order over 2 PDF files (total 18 pages) using our proposed method based on the LR structure over the baseline method using the LR structure in 20601 revision, which its accuracy of the reading-order is 0%. Until now, the footer issue and a part of multiple mini separated bounding box issue can be solved by using the MESH method. However, there are still three issues that cannot be solved, such as the table issue, formula issue, and the random multiple mini separated bounding boxes. But the detection of the table position and the recognition of the table structure are out of the scope in this paper, and there is needed another research. In the future, the tasks are chosen- how to detect the table position in the page and to extract the content of the table.

Keywords : document processing, reading order, observation method, layout recognition
Conference Title : ICDAR 2022 : International Conference on Document Analysis and Recognition
Conference Location : Jerusalem, Israel
Conference Dates : November 29-30, 2022