Computing Reviews
Association for Computing Machinery
ThinkLoud

Review

Digital weight watching: reconstruction of scanned documents

Marx M., Gielissen T. International Journal on Document Analysis and Recognition 14(2): 229-239, 2011. Type: Article

Date Reviewed: 07/25/11

In this paper, Marx and Gielissen construct a search system for parliamentary proceedings that have been scanned and processed via optical character recognition (OCR). This type of document has a large file size (in the targeted dataset, the average size was 16.5 MB per document). Therefore, users have to spend a lot of time downloading and going through documents that are highly ranked by search engines before determining which documents are actually relevant. The authors solved this problem by generating a summary of documents (snippets) from reconstructed PDF documents (transformed from the OCRed data); they significantly reduced the file size (to 1.5 percent of the original), leading to a reduction in download time.

The authors focus on rule-based architecture to extract information from OCRed data, and evaluate it in terms of physical points (reduction in file size and processing time) and economic cost. However, in addition to the physical points, they should have evaluated their proposed approach from the user's point of view. If the authors were to employ natural language processing or information retrieval techniques to generate informative snippets, they could improve user satisfaction.

Another notable point in this paper is that it contains several helpful pointers to information on making "governmental and/or political data easily accessible through the Internet." Researchers or developers who work for information systems departments in government offices can use the helpful references to get up-to-date information about this area.

Reviewer: Kazunari Sugiyama Review #: CR139275 (1201-0100)