Keyword Searching in Compressed Document Images

Yue Lu, Chew Lim Tan
Department of Computer Science, School of Computing
National University of Singapore, Singapore 117543
{luy,tancl}@comp.nus.edu.sg

A huge amount of document images are accessible in the Internet and digital libraries. We find that, most of them are packed in PDF files and are compressed using CCITT Group 4 standards for saving storage space and speeding up transmission. There is thus significant meaning to develop the methods of directly searching keywords from these documents. In this paper, we present a compressed pattern matching method for searching keywords from the CCITT Group 4 compressed document images, without explicit decompression.

According to the CCITT Group 4 standards, each coded position indicates that the current pixel color is different from its previous pixel, except for the next coded positions of the pass mode. In our work, we extract these changing elements from the compressed images directly. The changing elements are utilized to segment and bound the word objects, and are used for measuring the similarity of two word images. The connected components are labeled based on the line-by-line strategy according to the relative positions between the changing elements of the current coding line and the changing elements of the reference line. The connected components with small areas are eliminated as noise, while that larger than a threshold, which may be graphics or table regions, are ignored too. The word boxes are bounded by merging the remainder connected components according to their relative positions and sizes.

A two-stage matching strategy is constructed to measure the dissimilarity between the template image of the user's query word and the words extracted from document images. The first stage is a coarse-matching procedure, in which the distribution of changing elements is applied as feature. In the second stage, a weighted Hausdorff distance is employed to measure the dissimilarity between them, in which the contribution of different parts of the word object to the Hausdorff distance is not the same.

Experiments were conducted to verify the validity of the proposed approach. The document images are selected from the scanned books and students' theses that are provided by the Digital Library of the National University of Singapore. The document images are packed in the PDF files, and compressed by CCITT Group 4 standards. To evaluate the performance of the system, 95 images of scanned books and 324 images of scanned students' theses are included in the test. 40 keywords are selected to search their corresponding words from the document images. The system achieves an average precision ranging from 93.72% to 98.17% and an average recall ranging from 85.16% to 95.86% depending on different thresholds of the dissimilarity measurement. The experiments also found that the proposed algorithm is approximately 2.1 times faster than the traditional approach of decompression followed by word matching. This could be explained by the fact that the proposed approach avoids the pixel-level processing for analyzing the connected components and extracting word features, whereas the processing on image pixels in the traditional approach is quite time consuming.