Abstract

Arabic content on the Internet and other digital media is increasing exponentially, and the number of Arab users of these media has multiplied by more than 20 over the past five years. There is a real need to save allocated space for this content as well as allowing more efficient usage, searching, and retrieving information operations on this content. Using techniques borrowed from other languages or general data compression techniques, ignoring the proper features of Arabic has limited success in terms of compression ratio. In this paper, we present a hybrid technique that uses the linguistic features of Arabic language to improve the compression ratio of Arabic texts. This technique works in phases. In the first phase, the text file is split into four different files using a multilayer model-based approach. In the second phase, each one of these four files is compressed using the Burrows-Wheeler compression algorithm.