A NEW TEXT COMPRESSION TECHNIQUE BASED ON LA NGUAGE STRUCTURE

AKMAN, KI

Abstract

This paper describes a new data compression technique which utilises some of the common structural characteristics of languages. The proposed algorithm is designed to partition a word into its root and suffix(es), which are then replaced by shorter bit representations. The method uses three dictionaries in the form of binary search trees and one character array. The first two dictionaries are for roots, whereas the third one is for suffixes. The character array is used for both searching compressible words and coding incompressible words. The number of bits in representing a substring depends on the number of the entries in the dictionary in which the substring is found. The proposed algorithm is implemented in the Turkish language and tested using three different text groups with different lengths. The results indicate a compression of up to 47%.