ABSTRACT

ABSTRACT

Development of Word-Based Text Compression Algorithm For Indonesian Language Document

Ardiles Sinaga
Supervisor: Ir. Hertog Nugroho M.Eng., Ph.D
Co-Supervisor: Dr. Adiwijaya S.Si., M.Si.

Information technology is growing very rapidly, in particular for data handling. Data is a valuable asset for everyone, especially for larger companies with branches in several places. Data transmission from headquarters to branch offices make the company must provides good tools to do it. These companies also need tools that can be used to compress data to reduce their size.

The main idea of the word-based encoding is to extract each word of the source text, then it is checked whether containing capital letters or not. After that, it is checked if there is a symbol or number. The particle will be separated from the basic word using stemming algorithm. Symbols, numbers and affixes will be indexed in the basic dictionary. The basic word will also be checked whether it exists in the basic dictionary or not. If there not a match, then the word will be stored to the supplement dictionary.

The experiment was conducted on the text file with the size from about a 10.000 bytes up to 500.000 bytes and a code with length of bits is 16 bits. The result shows that the compression ratio of the proposed method is comparable with popular RAR application up to 200 kbyte, while its processing time is much better than the Reversed Sequence of Characters on LZW method.

Keywords: Data Compression, WB-LZW, Word-Base, Stemming, Tree, Basic Dictionary, Main Dictionary