Extending Huffman Coding for Multilingual Text Compression

Chi-Hung Chi, Chi-Kwun Kan, Kwok-Shing Cheng, Ling Wong
Department of Computer Science
The Chinese University of Hong Kong
Shatin, New Territories
Hong Kong
Email: chchi@cs.cuhk.hk

Traditional text compression algorithms such as Huffman and LZ variants are usually based on 8-bit characters sampling. However, under the unicode representation for multilingual information, the character set of each language such as Chinese and Japanese is consisted of a very number of distinct characters and thus 16-bit or 32 bit character sampling is needed. Consequently, when text compression algorithms based on 8-bit character sampling is applied to documents using 16-bit or 32 bit character sampling, very poor data compression ratio (average about 1.5) is obtained.

In this paper, we propose two new algorithms that are based on the 16-bit or 32 bit sampling character set and on the unique features of the languages with large number of distinct characters to improve data compression ratios for multilingual text documents significantly. We choose Chinese language using 16 bit character sampling (such as Big-5 or GB code) as the representative language in our study.

The first approach, called the Static Chinese Huffman Coding (Huffs), is to introduce the concept of a single Chinese character in the Huffman tree. Experimental results on our PH corpus showed that the improvement in compression ratio obtained by Huffs ranges from 20% to 29%.

The second approach, called the Dictionary-Based Chinese Huffman Coding (HuffDict), is to include the concept of Chinese words into the Huffman coding. A Chinese word \( y \) is defined as \( \alpha_1, \alpha_2, ..., \alpha_n \), where \( \alpha_i \in \text{single Chinese character set} \), \( 1 \leq i \leq n \), and \( n = \text{length of the Chinese word } y \). Word-based data compression for Chinese is not as straight forward because there is no "space" delimitation (as compared to English). Due to this word segmentation problem in Chinese, a standard Chinese dictionary is used to help segmenting Chinese words in a Chinese document. The segmentation of a Chinese document into "words" is equivalent to finding strings of Chinese characters in the document that are defined in the Chinese dictionary. Our experiments on the same PH corpus showed that HuffDict can improve the compression ratio by 19% to 72%, which is substantial.

(The complete paper can be obtained from Dr. Chi, Chi-Hung.)