要 旨

LZSS 圧縮アルゴリズムの日本語向け　改良に関する研究

久保 賢弘

近年では，携帯電話やパソコンの普及に伴い電子メールなどのメッセージ交換は日常的に使用されている．そういった中で，少しでも通信回線や中継サイトの負担を軽減することは非常に重要である．また，国内での転送データの多くは日本語文書で占められている．よって，日本語文書を効率的に圧縮できれば，回線や中継サイトの負担が減少し，低コストでメッセージ交換が可能になる．通常，データの圧縮には汎用圧縮ツールが用いられるが，圧縮対象となるデータが決まっていれば，そのデータの性質を利用してより効果的な圧縮が可能となる．

そこで本稿では，圧縮法として広く研究され，利用されている LZSS（Lempel-Ziv-Storer-Szymanski）法を改良して，圧縮の対象を日本語文書に限定し，日本語文書の特質を利用した手法を提案する．この圧縮手法では，圧縮に必要な文字列の位置情報や長さ情報を 1 バイト単位ではなく，2 バイト単位で扱うことにより，1 ビットずつ短い符号で表現することが可能となる．この手法で実際に日本語文書を圧縮した場合，従来の LZSS 法よりも高い圧縮率が得られることが確認できた．

キーワード　データ圧縮，日本語文書

Abstract

Improvement of LZSS algorithm for Japanese Text compression

Masahiro KUBO

In recent years, a huge number of e-mails are used by mobile telephones and PCs. In Japan, however, quite a number of them are Japanese text data. Though general-purpose compression tools are used generally, if the contents of the data is known customized compression tool would work better.

This paper proposes a improvement of LZSS (Lempel-Ziv-Storer-Szymanski) compression method for Japanese text. LZSS is a well-known universal compression algorithm, which is one of variations of LZ77 compression by Ziv and Lempel.

In proposed method, input data is dealt as a sequence of 2-byte characters instead of ordinary characters. The length of code for strings in the sliding window becomes shorter, because both the location and the length of strings are counted in 2-byte. This paper also shows the results of experiments. The proposed method can achieve better compression ratio than the original method.

key words Data Compression, Japanese Text