要 旨

日本語向け LZSS 圧縮アルゴリズムの圧縮効率向上に関する研究

永井 佐知

通常，データの圧縮には汎用ツールが用いられているが，圧縮対象の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になる．ここでは，日本語文書データを圧縮対象とした場合に高い圧縮率を実現するような圧縮方法について考察する．先行研究において，久保は，よく知られている LZSS 法の改良である日本語向け圧縮法を提案し，LZSS 法より約 2〜3%圧縮率を改善できる事を実験的に示した．久保の手法は LZSS 法と同じく，事前に圧縮対象内の文字の出現頻度等を調べる必要がない．伊藤らは，出現頻度の多い文字に短い符号を与える事による日本語向け圧縮法を提案している．この方法は文字の出現頻度を予め調べる必要がある．また，反復部分を省略するといった通常の圧縮技術が取り入れられていない．

本稿では，伊藤らの手法を参考にして久保の手法を改良し，日本語文書データに対してより高い圧縮率を得る手法を提案する．提案法では，2 バイト文字のうち，平仮名，カタカナと句読点を含む記号を合計 256 文字選んで短縮文字として 1 バイトで表現し，久保法よりも短い符号で表現する．従来の LZSS 法，久保法，提案手法を比較するために，久保が実験で用いた日本語テキストファイルを圧縮対象として圧縮率を測定した．実験により，提案法によって LZSS 法よりも 3〜4%，久保法よりも 1〜2%高い圧縮率を得られる事を示した．

キーワード　データ圧縮，日本語文書，LZSS 法

Abstract

Improvement of the Compression Rate of an LZSS-Based Algorithm for Japanese Text

Sachi NAGAI

Using the characteristics of data, more efficient compression than a universal compression tool can be achieved. In this thesis, we study a compression method with a better compression rate for Japanese text data. Kubo proposed an improved LZSS for Japanese text. He empirically showed that the compression rate of his method is 2∼3% better than the original LZSS. Ito proposed another compression method for Japanese text that gives a frequently appearing character a short code. One of the drawbacks of his technique is that it must examine the frequency of each character in advance. In this thesis, we improve the Kubo method by taking the method of Ito into account to archive higher compression rate for Japanese text data. We chooses 256 characters of hiragana's, katakana's and signs including the punctuation marks among double-byte characters as *short characters* and express each of them by one byte, which is shorter than the Kubo method. We conduct an experiment measuring the compression rate of the proposed method using the same Japanese text files as Kubo's experiment. The proposed method archived 3∼4% better compression rate than the original LZSS and 1∼2% better than the Kubo method.

key words data compression, Japanese text, LZSS