

テキストファイルのデータ圧縮

名列番号 9 氏名 大林 信

1. まえがき

データ圧縮にはさまざまな種類があり、それぞれが特徴を持っている。例えば、テキストデータの圧縮に適しているものや、画像データの圧縮に適しているもの、音楽のデータ圧縮に適しているものなどがある。それらは私たちの日常でよく使われているものが多く、また、最近ではコンピュータを扱う上では欠かせないものになっている。今後は、より効率のよい圧縮ができるものが求められるだろう。

2. 研究課題

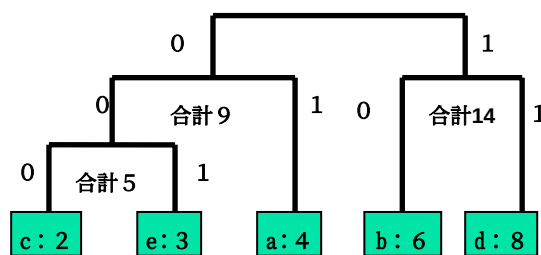
今回は、テキストデータの圧縮について圧縮方法のおもな種類の調査と、それらのプログラムを作成し、その効果を比較・検討する。

3. 研究方法

インターネットなどで圧縮方法の種類とその使用方法を調査し、そのいくつかについて自分で圧縮・展開のプログラムをC言語を用いて作成してみる。作成したプログラムが正常に動くかを確かめ、いくつかの条件でそれぞれの圧縮方法の効果を比較してみる。

4. 実験と考察

今回、実際にプログラムを作成した圧縮方法は連長符号化とハフマン符号化である。連長符号化は一番単純な圧縮法で、同じ記号がたくさん続く場合にしか圧縮効果がないので、主として画像データの圧縮に利用されている。ハフマン符号化は以下に示したような符号の木を用いた圧縮方法である。また、それらの比較には以下のような2つの条件を用いた。



記号	符号
a	01
b	10
c	000
d	11
e	001

※記号の横にある数字は出現回数

比較結果1

条件：同じ記号が連続で出現する場合(もとのデータのサイズ：3044byte)

結果：連長符号化 → 52byte
ハフマン符号化 → 740byte

連長符号化の方が圧縮率がいい

比較結果2

条件：ランダムに記号が出現する場合(元のデータのサイズ：2463byte)

結果：連長符号化 → 2460byte
ハフマン符号化 → 1521byte

連長符号化ではほとんど圧縮効果が見られず、ハフマン符号化の方が圧縮率がいい

5. まとめと今後の課題

圧縮方法にはさまざまなものがあるが、それぞれがどんなデータの圧縮に向いているかを理解して、目的にあったものを利用しなければならない。今回調査したほかの圧縮方法についても、一度、自分でプログラムを作成してみる。さらに、最近ではいくつかの圧縮方法を組み合わせているものもあり、そのことについても、今後、考えてみる。自分でプログラムを作成することによって、実際にどのような動作をしているかがよく理解できたのでよかった。