4U-05

二段階圧縮法の圧縮率の検討*

大塚真吾† 宮崎収兄‡

1 はじめに

コンピュータやインターネットの普及により、電子メールやホームページなど電子文書の持つ情報の重要性が高まっている。インターネットから文書ファイルをダウンロードする場合、通信コストやトラフィックの軽減のため何らかの可逆圧縮が行われる。圧縮されたテキストファイルに対して検索を行う場合、復号処理を伴うため高速化が困難である。この問題の解決策として、圧縮ファイルを直接検索する手法[1]や索引を用いた符号化法[4]などがある。

以前、我々は索引を用いた符号化法の1つである二段階圧縮法を提案した[2]。二段階圧縮法は索引を用いてテキストファイルを符号化し、そのファイルと索引を更に他の符号化法で圧縮を行っている。これにより圧縮率と検索時間が向上する。本稿では索引と2回目の圧縮に適した符号化法について検討を行う。

2 二段階圧縮法

二段階圧縮法は図1のように索引部と本体部とに分かれる。まず、圧縮対象となるファイルから単語を抽出し索引ファイルを作成する。次に、索引ファイルにある単語の位置情報を用いて圧縮対象となるファイルを符号化する（第一段階圧縮）。その際に符号語長以下の単語に対しては符号化を行わない。更にそのファイルを他の符号化法で圧縮を行う（第二段階圧縮）。また、索引ファイルも他の符号化法で圧縮を行う（索引圧縮）。

復号については符号化手順の逆を行えばよい。検索は索引部を復号してから行う。索引部は本体部に比べ非常に小さいため、復号時間は小さい。

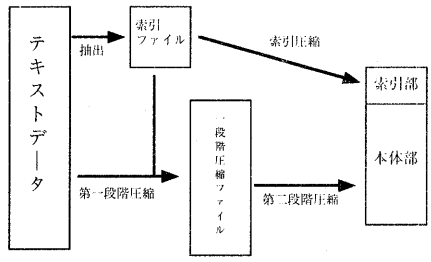

図 1: 圧縮の概要

3 索引圧縮、第二段階圧縮に適した符号化法

ファイルを圧縮する際に用いる符号化法は大きく分けて、

1. 単語や文字などを用いて符号化を行う
2. ビット単位に符号化を行う

の2つに分けられる。1はアルファベットの頻度に基づいて符号語を生成する「ハフマン符号」などがある。一方、2についてはビット列を局所辞書に登録し、これを用いて符号化を行う「LZ77符号」やビット列を大域辞書に登録する「LZ78符号」等がある。一般的に2の方が圧縮率が良く圧縮ソフトなどに利用されている。しかし、テキストだけに限定すると1の一種である文献[4]のarith_coderなどは高い圧縮率を得ることができる。

索引ファイルを圧縮する場合、ファイル自体はすでに重複単語を取り除かれている状態にある。また、第二段階圧縮を行うファイルは索引ファイルを用いて単語を符号化しているので、ファイル中に単語がほとんど存在しない。したがって、どちらの圧縮もビット単位符号化の方が圧縮率が良いと考えられる。

---

*A study on compression of two-stage compression

†千葉工業大学工学研究科情報工学専攻 otsuka@mz.cs.it-chiba.ac.jp

‡千葉工業大学工学部情報工学科 miyazaki@cs.it-chiba.ac.jp

表 1: 英語記事の結果

圧縮ソフト	圧縮率	二段階圧縮法に適用
CAB	31.63%	33.19%
RAR	34.27%	34.86%
LHA	39.84%	37.48%
ZIP	41.23%	38.21%
arith_coder	30.09%	40.58%

4,812,730バイト

表 2: 日本語記事の結果

圧縮ソフト	圧縮率	二段階圧縮法に適用
CAB	20.10%	21.80%
RAR	21.41%	22.76%
LHA	25.78%	25.37%
ZIP	48.63%	42.64%

2,924,220バイト

4 実験と評価

4.1 実験環境

一般に使われている圧縮ソフトを用いて実験を行った。使用したソフトは CAB,RAR,LHA,ZIP である。CAB は Microsoft の標準のキャビネット形式であり、LHA,ZIP は一般的に広く使用されているソフトである。RAR は比較的新しい圧縮形式で LHA や ZIP より圧縮率が良いとされている。これらのソフトは単語を使った符号化を行っておらず索引圧縮と第二段階圧縮に適していると思われる。

また、比較として単語を使って符号化を行う arith_coder でも実験を行った。

実験は英語で書かれた雑誌記事のファイルと日本語で書かれた新聞記事を対象に行った。

4.2 実験結果

実験結果について表 1,2 に示す。表中の左の数字はファイルをその圧縮ソフトで圧縮した場合の圧縮率である。右の数字は二段階圧縮法の第二段階圧縮と索引圧縮にそのソフトを用いた場合の圧縮率である。二段階圧縮法の結果では索引のサイズも含んでいる。

英語テキストでの結果から LHA と ZIP はファイルをそのまま圧縮するより、二段階圧縮法を用いた方が良い結果となった。また、その他のソフトでは圧縮率が若干悪くなるものの良好な結果が得られた。また、単語を使って符号化を行っている arith_coder は圧縮率がかなり良いが、二段階圧縮法に用いると圧縮率が悪くなる。

日本語テキストの実験では arith_coder が日本語に対応していないため実験を行わなかった。その他の実験結果は英語テキストと同様な傾向が見られた。

4.3 まとめ

二段階圧縮法の索引圧縮と第二段階圧縮に既存の圧縮ソフトを利用すると圧縮率はそのソフトの性能に依存している事がわかる。また、英語テキストでは arith_coder が二段階圧縮法に適していないこともわかる。

5 おわりに

本稿では圧縮率と検索効率を考慮した二段階圧縮法の索引圧縮と第二段階圧縮に適した符号化法について検討を行った。二段階圧縮法の性質上、索引圧縮と第二段階圧縮には単語単位に符号化を行わない符号化法が適している。

参考文献

[1] 松本光崇, 角田達彦, 松本裕治. 圧縮ファイルへの直接検索を可能にする符号化法の提案. 電子情報通信学会論文誌, Vol. J79-A, pp. 41–48, 1996.

[2] 大塚真吾, 宮崎収兄. 高速検索を可能とする日本語テキストの二段階圧縮法. *DEWS2000*, 2000.

[3] 定兼邦彦, 今井浩. 転置ファイルおよび接尾辞配列の効率的圧縮法. 情報処理学会論文誌データベース, Vol. 40-No. SIG 8(TOD 4), pp. 85–94, 1999.

[4] J. Zobel and A. Moffat. Adding compression to a full-text retrieval system. *Software-Practice and Experience*, Vol. 25-8, pp. 891–903, 1995.