I-14

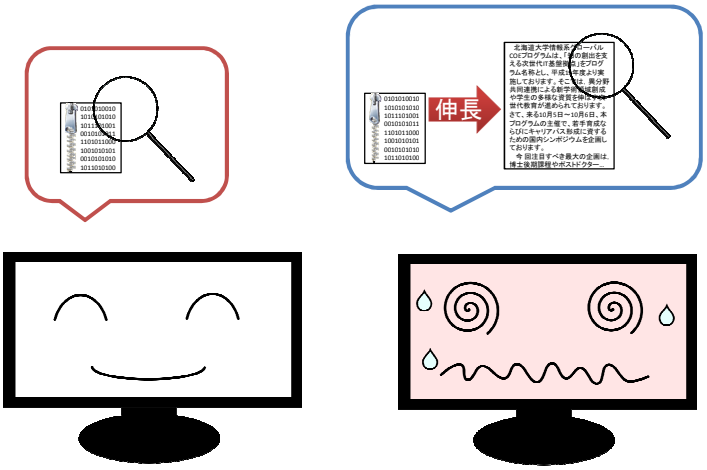
On Performance of VF Codes

吉田諭史*，喜田拓也*

*北海道大学　大学院情報科学研究科　コンピュータサイエンス専攻

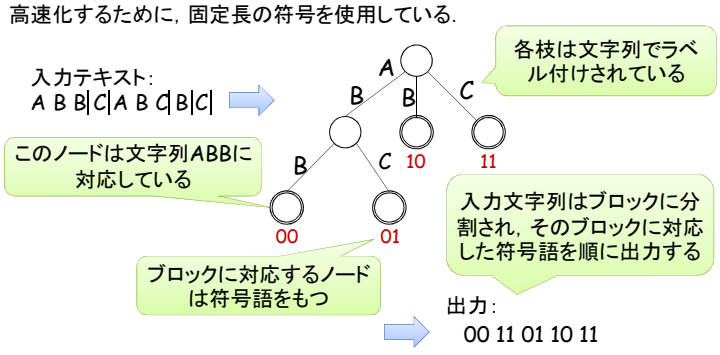
まえがき

巨大な記憶装置が取り扱えるようになった現在では，莫大なデータから効率のよい情報検索する技術が重要視されている．莫大なデータは保存領域を削減するため，圧縮された形式で保存されることが多い．しかし，そのようなデータを利用するときには伸長する必要がある．莫大な圧縮データを伸長するのは時間がかかり煩わしいため，結局のところ，利用されずに放置されることになる．そのような背景から，圧縮データを伸長せずに検索する技術が1990年代半ばから議論されてきた．現在では，圧縮の方法をうまく選ぶことで，効率のよい検索が行えることがわかっているが，このような圧縮方法は，gzipやbzip2に比べて圧縮率が劣る．そこで，本研究では，実用的な圧縮性能と検索の容易さとを両立することを目指す．



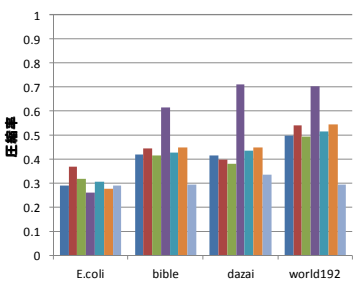
イノベーティブポイント

データ圧縮では，通常，圧縮率が重要視される．このため，高い圧縮率を達成する可変長符号が現在の研究の主流である．しかし，本研究では，情報検索を高速化するために，固定長の符号を使用している．

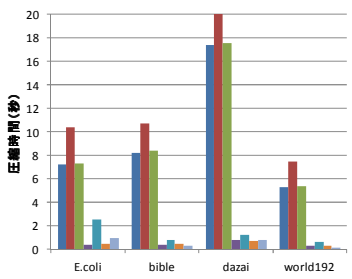


固定長符号は，可変長符号に比べて圧縮率が劣る．そこで，本ポスターでは，固定長符号と他の圧縮方法との組み合わせの圧縮率と圧縮速度を調査した．

- STVF符号単体
- STVF符号 + Huffman符号 (STVF + Huf)
- STVF符号 + Range coder (STVF + RC)
- Tunstall符号単体
- Tunstall符号 + Huffman符号 (Tunstall + Huf)
- Tunstall符号 + Range coder (Tunstall + RC)
- gzip



圧縮率の比較



圧縮速度の比較

圧縮方法を組み合わせることで，圧縮率が良くなった．また，Range coderを組み合わせた場合は，圧縮速度はほとんど変わらなかった．

社会への有効性

本研究では，圧縮された文書を伸長せずに検索を行う．そのため，巨大な文書を，保存する領域を削減しながら活用することができる．また，2000年以降，情報検索の効率が良い圧縮の方法がいくつか提案されてるが，そのほとんどは可変長符号を採用している．本研究では，それらとは異なり，固定長符号を採用している．固定長符号は，すべての符号語が等しい長さをもつので，圧縮データが取り扱いやすい．そのため，可変長符号と比べて，高速な検索ができる．

固定長符号

- すべての符号が同じ長さ
- 先頭からでなくても復号できる
- 符号語を容易に切り出せる
- 処理が単純

a　b　c　a　b

00　01　00

可変長符号

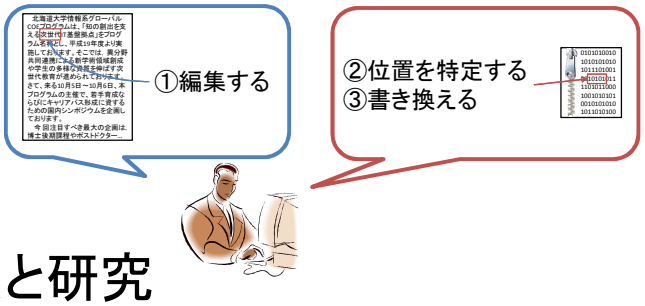
- 符号の長さはバラバラ
- 先頭から順に読まないと復号できない
- 符号語の切り出し操作が大変
- 処理が複雑になりがち

a　b　c　a　b

0　01　001　0　01

未来社会へ向けた新技術

・圧縮していないデータと同じようにアクセス可能な圧縮ファイルシステム



私と研究

・本研究における貢献度

本研究では，Huffman符号およびRange coderの実装，組合せ手法の実装，実験を行った．Huffman符号やRange coderは，広く用いられている手法ではあるが，インターネット上で公開されているプログラムでは，入力に制約があるため，独自に実装を行った．

・今後の課題

今後も，固定長符号の圧縮率を向上させる手法について研究を続けていく．また，圧縮された文書高速な情報検索の手法の開発も行う．

研究業績

[1] Satoshi Yoshida and Takuya Kida: “An Efficient Algorithm for Almost Instantaneous VF Code Using Multiplexed Parse Tree,” Proc. of Data Compression Conference 2010 (DCC 2010), pp.219-228, IEEE, Snowbird, Utah, USA, 2010.

[2] Takashi Uemura, Satoshi Yoshida, Takuya Kida, Tatsuya Asai, and Seishi Okamoto: “Training Parse Trees for Efficient VF Coding,” Proc. of the 17th String Processing and Information Retrieval (SPIRE 2010), pp. 179-184, Springer, Lecture Notes in Computer Science, vol. 6393, Los Cabos, Mexico, October 2010.

[3] Satoshi Yoshida and Takuya Kida: “A Combination of Variable-length-to-Fixed-length Coding with Arithmetic Coding for Efficient Compression and Pattern Matching,” 5th Workshop on Compression, Text, and Algorithms, Los Cabos, Mexico, October 2010.

[4] Satoshi Yoshida and Takuya Kida: “On Performance of Compressed Pattern Matching on VF Codes,” Proc. of Data Compression Conference 2011 (DCC 2011), p. 486, Snowbird, Utah, USA, March 2011.

[5] Satoshi Yoshida, Takashi Uemura, Takuya Kida, Tatsuya Asai, and Seishi Okamoto: “Improving Parse Trees for Efficient Variable-to-Fixed Length Codes,” IPSJ Journal, 2011 (accepted for publication).