1C－8

テキスト圧縮を用いた全文検索高速化の一手法

井口　博彰†　黒須　康雄†　横山　佳弘†　藤縄　雅章††
†(株)日立製作所　マイクロエレクトロニクス機器開発研究所
††(株)日立製作所　ストレージシステム事業部

1．はじめに

近年，テキストデータベースの飛躍的な増加に伴い，インデキシング作業を必要としない，全文検索手法への注目が高まっている[1]。これまで様々な全文検索高速化手法へのアプローチが提案されているが，テキスト蓄積媒体からの読出し速度が，検索処理速度のボトルネックとなっている[2]。

これに対し，文献[3]では，Huffman符号を用いた圧縮テキストに対する直接文字列パターン照合手法が提案され，英文テキストに対し圧縮比率60%，検索処理時間比率70%が得られている。しかし，上記手法の問題点として，可変符号長の圧縮コード処理に伴う検索処理量の増加がある。

本報告では，圧縮コード長を2バイトに統一する，検索処理量の増加を伴わない高速な圧縮テキスト直接照合手法の提案を行う。

2．文字列コード化圧縮手法

高速検索を実現する圧縮手法として，2バイト統一長圧縮コードを用い，符号長を統一する。符号化手法を以下に示す。2バイト圧縮コードの領域は，2^{16}＝ 65,536コード存在する。日本語テキストのコード表記法であるMS-DOS™*テキストコードあるいはＪＩＳ Ｘ 4004テキストコードにおいて，文字キャラクタは約9,000種類である。従って，文字キャラクタ以外の領域，約5万6千コードに熟語，接続詞等を割り当て，テキストを圧縮する手法を提案する。本手法を，文字列コード化圧縮手法とする。

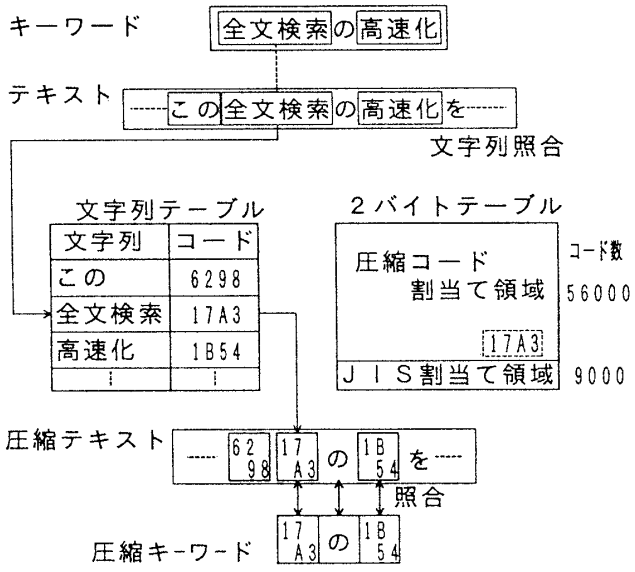

図1　文字列コード化圧縮手法

文字列コード化圧縮手法を，図1に示す。圧縮には，文字列テーブルと文字列照合手法を用いる。すなわち，テーブルに登録された文字列がテキストデータと照合した場合，テキストデータ中の照合文字列と2バイト圧縮コードを置き換えることによりテキスト圧縮を行う。

3．圧縮コード直接照合処理による検索高速化

上記文字列コード化圧縮手法の利点は，ワード比較による圧縮テキスト直接照合処理にある。すなわち，2バイト圧縮コードは原テキストの文字列に1対1に対応し，テキストとキーワードは同じ圧縮コードに変換される。従って，図1に示すごとく，検索実行時にキーワードを圧縮コード変換することにより，圧縮テキストと圧縮キーワードのワード単位での直接照合が可

Fast Full-Text Retrieval Method by Using Data Compression
Hiroaki IGUCHI, Yasuo KUROSU, Yoshihiro YOKOYAMA, Masaaki FUJINAWA
ＨＩＴＡＣＨＩ, Ltd.

*MS-DOSは米国マイクロソフト社の登録商標です

能となる。本手法では，検索時にビットシフト等の特殊な処理を必要とせず，検索処理量は増加しない。よって，テキスト圧縮比率と検索処理時間比率は等しい値が得られる。

4．日本語に対する文字列テーブル構成

上記文字列コード化圧縮手法の圧縮効率は，文字列テーブルの構成に大きく依存する。次に，日本語テキストを対象とした文字列テーブル構成に関する検討結果を述べる。表１に，日本語テキスト圧縮に適切なテーブル構成を示す。

（１）静的辞書法

日常一般的に用いられる基本熟語は２万語程度である。この特徴を利用し，あらかじめ定めた熟語を文字列テーブルに登録しておく。上記手法を，静的辞書法と呼ぶ。

（２）文字連接法

ひらがな或は英文字２文字の連接を文字列テーブルに登録し，１つの圧縮コードに置き換える。この手法を文字連接法と呼ぶ。

（３）登録辞書法

カタカナの圧縮コード変換手法として，文字種が変化する点を用い単語を切り出し，これを文字列テーブルに順次登録していく手法を提案する。上記手法を登録辞書法と呼ぶ。

5．テキスト圧縮実験結果とその検討

上記提案を行った圧縮手法について，実際の日本語テキストに対し実験による評価を行い，その効果を確認した。図２に，文字連接法，国語辞典のみ用いた静的圧縮法と，本手法との比較実験結果を示す。

実験結果より，文字連接法，静的圧縮手法に比較し，提案圧縮手法では全ての文字種に対し良好な圧縮が得られている。また原テキストデータに対する圧縮テキストデータの比率は，62％となった。上記結果より，日本語テキスト圧縮に対する提案圧縮手法の有効性が証明された。

表１　日本語テキストの特徴

特徴 / 文字種	単語の区切り	表記の変化	文字ｷｬﾗｸﾀ数	適用辞書	登録語数
漢　字	不明瞭	少ない	6353/0 (注)	静的辞書	28,000
ひらがな	不明瞭	多い（活用形）	83/0 (注)	静的辞書＋文字連接	9,200 +6,889
カタカナ	明　瞭	多い（表記のゆれ）	86/58 (注)	登録辞書	11,000
英文字	明　瞭	多い（活用形）	52/52 (注)	文字連接	1,352
数　字	明　瞭	無い	10/10 (注)	――――	9,058

（注）2ﾊﾞｲﾄｺｰﾄﾞ/1ﾊﾞｲﾄｺｰﾄﾞ

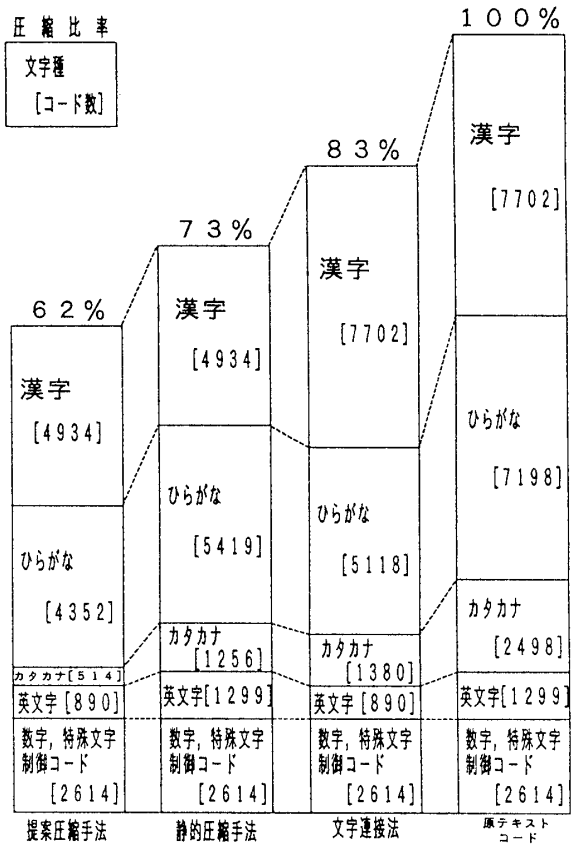

図２　圧縮手法比較実験結果

6．結論

テキスト圧縮を用いた新しい観点からのアプローチによる，全文検索高速化手法を提案した。更に，日本語テキストに対し，文字種の特徴を考慮した圧縮を用いることにより，62％の圧縮比率を得た。従って，日本語テキスト検索処理時間は，62％に短縮される。

［参考文献］
［１］松尾，神尾：日本における新聞記事データベースの現状と今後の動向，情報管理，Vol.35,No.10,pp.871-883(1993.1)
［２］加藤，他：大規模文書情報ｼｽﾃﾑ用ﾃｷｽﾄｻｰﾁﾏｼﾝの研究，情処研報，Vol.89,No.66,pp.14.6.1-14.6.8(1989.7)
［３］深町，篠原：圧縮データのための高速文字列パターン照合技法，情処学会第43回全大予稿，，分冊４，pp.83-84(1991)