テキストファイルにおける圧縮率と検索効率の向上

6P－5

大塚真吾　高谷健文　宮崎収兄
千葉工業大学　工学部　情報工学科

1　はじめに

近年、コンピューターやワープロの普及によって大量の文書が電子化されている。また、電子メールやＣＤ－ＲＯＭの普及によって、これらの情報が容易に利用可能になってきた。これに伴い、このような情報を有効に利用するための検索技術へのニーズが高まっている。このような動向の１つとして従来の情報検索システムに対して、使いやすく柔軟な検索機能を持った情報検索システムとしてのフルテキストサーチシステム（全文検索システム）や、情報を効率良く格納するための圧縮技術の有効性が認識されている。本稿では、今まで個々に発展してきた圧縮技術と検索技術を統合させることで、テキストファイルにおける圧縮率と検索効率を向上する方法を提案する。

2　従来のテキスト検索方式の問題

2.1　フルテキストサーチ方式

・一般的に日本語検索の場合、検索する内容のほとんどが、漢字、カタカナ、英語の単語である。それに対し、フルテキストサーチでは検索する時、助詞などの「ひらがな」も照合しているので、その分検索時間が増えてしまう。
・テキストを圧縮すると検索が困難なので基本的にそれ以上ファイルの容量は小さくならない。

2.2　索引による検索

・１つの例として、文書データファイルから索引ファイルを作り、文書データファイルを圧縮する。この方法では、索引ファイルを新しく作るために容量が増える。また、索引ファイルを圧縮する方法では検索時間に読み込みと文字列照合時間が含まれる。

3　索引圧縮ファイル方式

従来の方式ではそれぞれ問題点があるがその中でも我々は、フルテキストサーチで、助詞などを照合してしまい、その分検索時間がかかる点や索引による検索で、索引ファイルを作成する事で記憶容量が増えてしまう点に着目した。そこで、この２点を改善するために索引ファイルを使い文書データファイルを圧縮し更にそのファイルを他の圧縮方法で圧縮する二段階圧縮方式を提案する。索引ファイルを圧縮にも用いるので索引圧縮ファイル方式と呼ぶ。

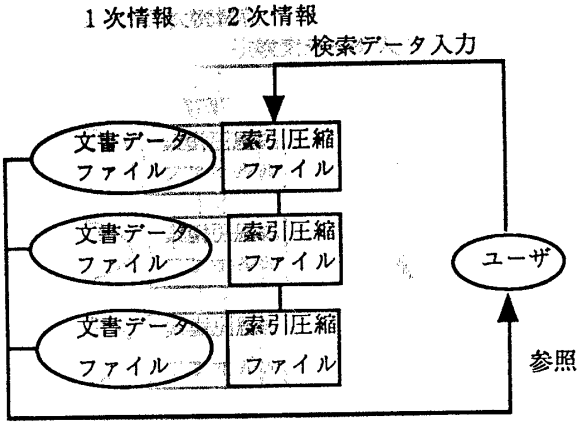


図１：索引圧縮ファイル方式の概要図

3.1　アルゴリズム

・圧縮方法
１索引圧縮ファイルを作成する。
（１）文書データファイルから単語を抜き出し索引圧縮ファイルに格納し、その単語があった場所に通し番号を付ける。
（２）同じ単語をまとめる。
（３）頻度順に並べる。
２索引圧縮ファイルで文書データを圧縮する。
・文書データファイルの通し番号を頻度順の通し番号に直す。
３文書データファイルを他の方法で圧縮する。

・復元方法
１文書データファイルを上記３の方法で復元する。
２索引圧縮ファイルを使い復元する。

3.2　単語の定義について

文書データファイルから単語を抽出する場合、構文や文脈から単語を抜き出す方式と機械的に単語を抜き出す方式があるが、今回はプログラムの作成が比較的容易で性能の良い後者の方式を採用した。

Integration of Compression and Retrieval Methods for Text Files.

Shingo Otsuka , Takefumi Takaya , Nobuyoshi Miyazaki
Department of Computer Science, Chiba Institute of Technology
2-17-1 Tsudanuma Narashino Chiba 275 Japan

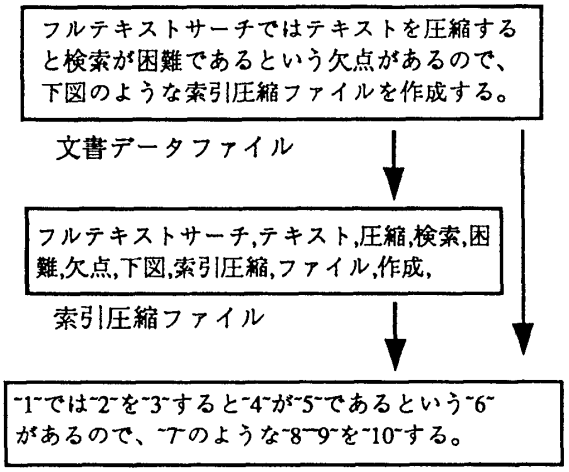
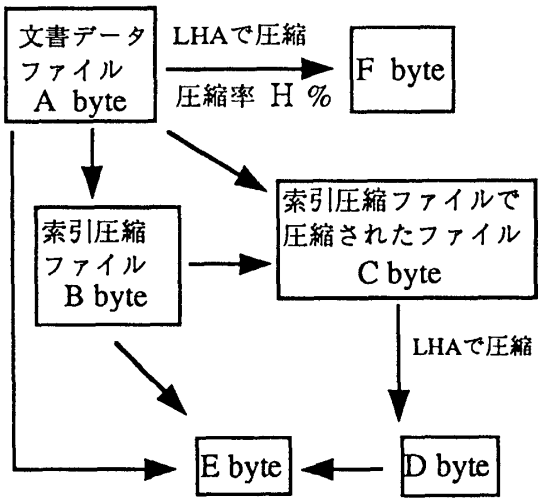


図２：プログラム処理例



圧縮率 G%＝E／A

図３：プログラム処理の概要図

表１：様々なファイルの処理結果

ファイル名	A	B	C	D	E	F	G	H
LHAの説明	16142	2280	14091	4413	6693	5731	41.5	34.9
小説	8678	1082	8746	4523	5605	4414	64.6	50.7
新聞	7618	3035	7319	2313	5348	4270	70.2	56.1
社説	5073	1831	5064	2843	4674	3139	92.1	61.9
ニュース	8583	3676	7833	4059	7735	5071	90.1	59.1
薬害エイズ	8251	1676	8066	4098	5774	4317	70	52.3
経済白書	52589	9865	50450	21297	31162	24725	59.3	47
建築基準法	153797	16050	142590	40307	56357	45338	36.6	29.5

・カタカナ、アルファベットは単語とする。
・１文字の漢字は主語、目的語（次に「は「が」「の」「を」がくる）は単語とする。
・２文字以上の漢字は形容動詞（次に「な」がくる）以外単語とする。
・漢字、カタカナ、アルファベットが連続する場合、それぞれ分けて登録する。

4　実験・評価

二段階圧縮をする際にＬＨＡを使い実験を行った。実験用のテキストデータとして、インターネットなどからの記事を利用した。

・圧縮率については表１の様な結果が得られた。
・文書データファイルを用いて二段階圧縮すれば文書データファイルをLHAで圧縮したより圧縮率が０～５０％向上する。
・文書の量が増加すると圧縮率が上がる。
・索引圧縮ファイル中の単語の半分以上が低頻度単語であるという結果になった。
・文書データファイルをそのままＬＨＡで圧縮するより １０％～１５％増加する。

5　まとめと今後の課題

文書データファイルから索引圧縮ファイルを作成し、そのファイルを使って文書データファイルを圧縮する手法を提案した。また、そのプロトタイプを作成し、実験を行い、索引圧縮ファイル方式の有効性を示した。今後の課題として以下のような点が上げられる。

・索引圧縮ファイルを実際にデータベースに乗せた場合の検索効率
．第１段階の圧縮率についてプログラムに改善の余地がある。

参考文献

[1] 程、桧、池田：“順次インデックスに対する差分圧縮法の具体的提案”情報処理学会論文誌、Vol.36 No9
[2] 菊池忠一：“日本語文書用高速全文検索の一手法”電子情報通信学会論文誌Vol．J 75-D-ＩNo 9