日本語文書に対する新しい索引検索方式
4F－3　— 試作・実験および評価 —

稲葉 光昭　野口 直彦　菅野 祐司　倉知 一晃
松下電器産業（株） マルチメディアシステム研究所

1 はじめに

著者は，単語辞書を援用することにより，索引量が小さく，高速な検索を可能にする新しい索引検索方式 [1] を考案した．本稿では，考案した原理に基づいて試作した実験システムの構成と，その実験システムを用いて新聞データを対象におこなった，索引量・検索速度の評価実験の結果について報告をする．

2 試作システム

試作した実験システムの構成を図 1 に示す．索引作成ツールは，検索対象文書から文書を読み込んで，論文 [1] で述べた方法により辞書にある単語を切りだし，その出現位置とともに索引ファイルに記録する．索引検索ツールでは作成された索引ファイルのみを用い，やはり論文 [1] で述べたアルゴリズムによって，検索条件として与えられた文字列が出現する文書中の位置をすべて求め，結果として出力する．

本方式による索引ファイルは，基本的には辞書中の単語と，各単語の検索対象文書中での出現位置との対応を記録した転置ファイルである．ただ，索引検索の高速化と索引量の圧縮のために，単語を検索対象文書中での出現頻度によって高頻度語と低頻度語に分類し，別々の索引構造を持たせている．また，いくつかの副次的なテーブル類がある．索引ファイルについての詳細は，次節で述べる．

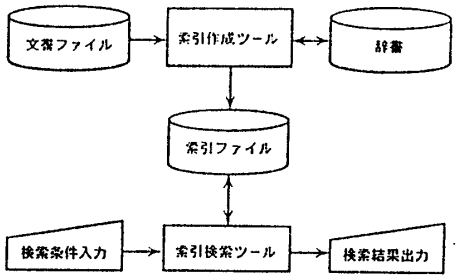

図 1: システムの構成

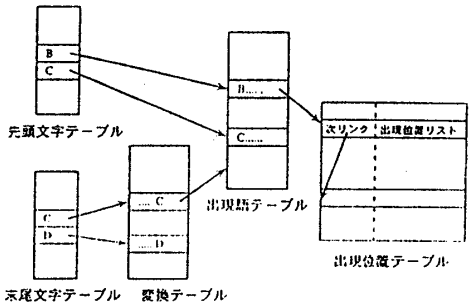

図 2: 索引ファイル

2.1 索引ファイル

図 2 に高頻度語に関する索引ファイル構成の概略を示す．索引検索の際，検索文字列の接頭／接尾延長語を高速に求める必要があるために，出現単語を辞書順にソートして格納する出現単語テーブル，検索対象データ中における当該単語の出現位置を格納する出現位置テーブル，出現単語の先頭文字／末尾文字をキーとする出現単語テーブルへのポインタである先頭文字テーブル／末尾文字テーブル／変換テーブルを持つ．

低頻度語に関しては，いくつかの語をグループ化して出現位置を記録する．このため，先頭文字テーブル／末尾文字テーブルに対して別々の出現単語テーブル／出現位置テーブルが必要になる．

また，検索文字列に対して両延長語を高速に求めるための情報として，１文字テーブルを持つ．１文字テーブルには，ある文字がどの出現単語の何文字目に現れるかという情報が格納されている．

New indices for Japanese text
-Their implementation, experiments and evaluation-
Mitsuaki Inaba, Naohiko Noguchi,
Yuji Kanno, Kazuaki Kurachi
Matsushita Electric Industrial, Co., Ltd.
5-15, 4-chome, Higashi-Shinagawa,
Shinagawa-ku, Tokyo 140 Japan

3 実験および実験結果

3.1 実験条件

検索対象として，電子化された朝日新聞の本文データを用いた．検索対象データの規模は実験により異なるが，最大で２年分 (約 170MB) である．

実験には Solbourne 社製 ワークステーション Sol5/800C(CPU:Sparc, 主記憶 128MB) を使用し，検索時間は CPU time にて計測した．

3.2 索引量

索引量は，高頻度語／低頻度語の分割の仕方に依存する．2.1 節で述べたように，低頻度語についてはグルーピングを行なっており，出現位置情報は二重に記録される．従って，低頻度語の割合を大きくすると，低頻度語の位置情報が全体の索引量に悪影響を及ぼす．一方，極端に高頻度語を増やすと，出現位置テーブルのヘッダ部と出現単語テーブルの要素数が増大して全体の索引量に悪影響を及ぼす．

高頻度語／低頻度語の分割の仕方を変化させたときの索引量の変化をグラフにしたのが図 3 である．図 3 の横軸は，高頻度語として選んだ語の数が全出現単語数に対

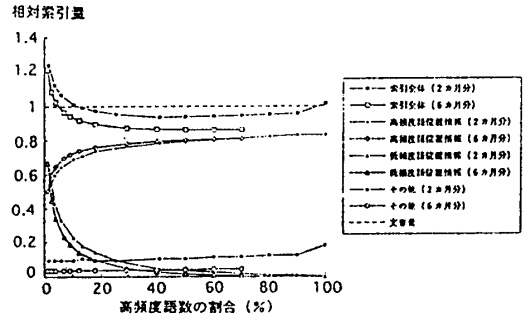

図 3: 高頻度語の割合と索引量の関係

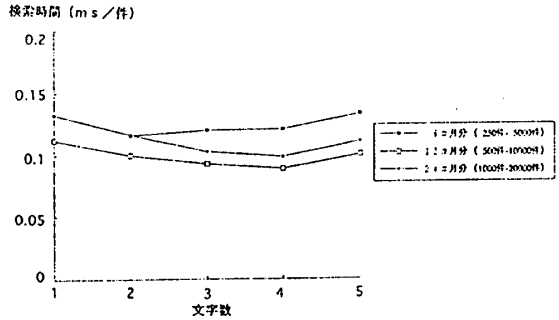

図 5: 文字数と 1 件あたり平均検索速度の関係

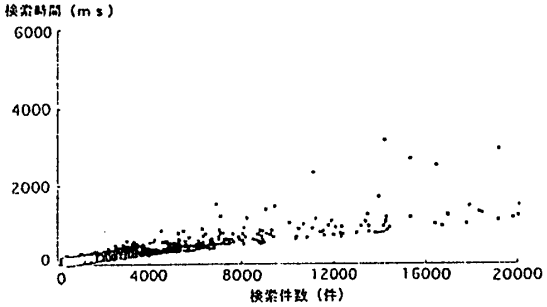

図 4: 検索件数と検索時間の関係（辞書単語）

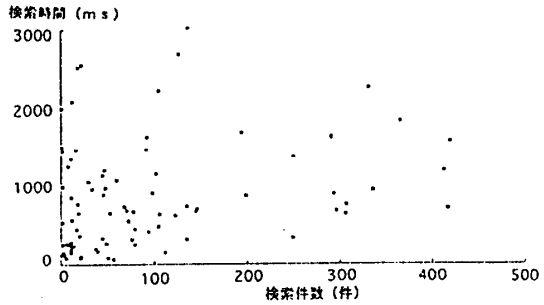

図 6: 検索件数と検索時間の関係（複合語）

して占める割合，縦軸は，原文書量に対する索引量の比である．高頻度語の割合を増加させると，索引量は漸減し，2 カ月分では高頻度語の割合 13 %前後，6 カ月分では，高頻度語の割合 5 %前後で索引量が原データ量を下回ることがわかる．

3.3 検索速度

3.3.1 辞書登録単語による検索

まず，検索文字列が辞書に登録されている単語である場合について，検索時間を測定した．実験には，6 カ月分，12 カ月分，24 カ月分の各データセットについて，切り出された全異なり単語数の 40 %を高頻度語とする方法で作成した索引ファイルを用いた．検索文字列としては，6 カ月分のデータから切り出された全異なり語を文字数別に分類したものを用い，検索文字列の文字数毎に検索時間を測定した．

図 4 は，24 カ月分のデータに対して作成した索引を用いて，3 文字語を検索した時の検索時間を各語の検索件数に対してプロットしたものである．検索結果の件数と検索時間はほぼ比例関係にあることがわかる．この傾向は，データ量，検索文字列長によらず，現れた．

検索の前処理時間が相対的に大きくなる，検索件数の少ない部分を除き，文字数別／文書量別に検索結果 1 件あたりの検索時間の平均を求め，グラフにしたものが，図 5 である．グラフからわかる通り，検索速度は文字数や文書量には依存していない．

3.3.2 複合語による検索

通常の検索文字列としては，いくつかの単語の組合せである複合語が与えられる場合が多いことが予想される．そこで，次に複合語を検索文字列として与えた場合の検索速度について実験を行なった．検索対象として 12 カ月分のデータを用い，検索文字列として，単語 2 語からなる複合語で，かつ，辞書の単語になっていないものを新聞データから任意に 100 語抽出したものを用いた．

図 6 は，これらの複合語に対する検索件数と検索時間の関係をグラフにしたものである．検索時間のばらつきが大きく，定性的な特徴は見られない．辞書単語の場合との定量的な比較は難しいが，結果 1 件あたりの検索時間は数十倍以上かかっていると考えられる．

4 考察

索引量に関しては，極大元のみを切り出す単語切りだし手法 [1] と，低頻度語のグループ化によって，原文書データ量以下に抑えられることを確認した．

検索速度に関しては，辞書登録単語による検索では，検索結果件数のみに依存し，検索対象データ量には依存しないということを確認した．例えば 2 年分 (約 65 万文字) 程度の文書に対して，500 件程度の結果を返す検索条件文字列に対しては，10 数億文字／秒の速度が出ていると考えられる．

5 おわりに

著者等は，検索速度が速く，索引の小さな，新しい索引検索方式を開発し，新聞記事データを用いた評価実験によりその有効性を確認した．

今後の課題として，実用化に向けて以下のような改良を行なっていく．

- 索引検索ツールのインプリメントの最適化
- 単語辞書の改良による複合語検索の高速化
- 索引作成時間の短縮

謝辞

当研究に関して，評価実験用に新聞記事の使用を許可してくださった朝日新聞社ニューメディア本部の方々に感謝致します．

参考文献

[1] 倉知，野口，菅野，稲葉: 日本語文書に対する新しい索引検索方式 −索引作成と検索の原理−，情報処理学会全国大会論文誌，4F-2 (1995).