7H－5

文字連接を用いたフルテキスト検索の高速化

宮原末治　　小橋史彦
（ＮＴＴ　ヒューマンインタフェース研究所）

１．まえがき　　多量に発生する情報を、事前加工なしにフルテキストの形で蓄積し、情報が必要になった時点で、自然言語を用いて検索するフルテキスト形データベース検索システムについて検討している[1)]。この中で、検索は「問い合わせ」と「検索結果からの適合文の選択」とを繰り返す仕事となり、これを効率よく行うために処理の高速化が要求され、ハードウェアで実現する試みがある[2)]。

本報告ではソフトウェアによってテキスト検索を高速に行う方法として、文字連接情報を索引に用いた検索法の検討結果について述べる。

２．外部条件と処理内容　　テキスト検索を行うためのハードウェアにはＥＬＩＳを用いた[3)]。ＥＬＩＳはマイクロコード長が６４ビットのマイクロプログラム制御で動作する装置であり、マシンサイクルは１２０nsecである。

この装置にテキストとして新聞記事を登載し、文を単位とした検索を行うことにした。検索の処理は、フルテキスト検索と同等の処理を実現することを狙いに、下記のような７つの処理ステップで構成した。

① 前処理（索引作成とテキストのリスト形式への変換）
② オペレータによる問い合わせ（検索文）の入力
③ 入力文からの単語（キー単語）抽出とシソーラスによる類語展開（これらの単語を総称して検索単語と呼ぶ）
④ 索引を用いた候補文の抽出
⑤ 候補文サーチによる検索単語の種類数と出現数との検出
⑥ 候補文の適合性の計算（キー単語ごとの種類数の総和に検索単語の出現数を加えた値）とその順位付け[1)]
⑦ 優先付けされた検索結果の表示・出力

この処理の中で①の索引付けの方法を変更した場合に必要となるメモリ容量と ④,⑤,⑥の処理を合わせた処理速度との関係、およびその時のフルテキストサーチの検索結果との差分を調査比較することとした。

３．実験と評価　　高速化の方法としては、語句が出現する文番号を索引として予め用意しておき、検索の際には検索文の検索単語に出現する語句から候補文を求め、その候補文に対してのみテキストサーチの処理を施し、検索単語と同じ単語が存在するか否かを検査する方法を採用した。

表１．索引の付与方法とメモリ容量・候補文数の関係

方法	索引の付与方法	平均候補文数	ﾒﾓﾘ容量*（Kﾊﾞｲﾄ）	索引の種類数
イ	テキストからの文字索引	６２	６５５	2,０８４
ロ	テキストからの連接索引	4	６７２	２２,０２５
ハ	解析単語からの連接索引	5	４８７	１３,９５３

＊ テキスト(被検索文)の大きさは約６０５Kﾊﾞｲﾄ（約 5,000文）

＊ A Study of Hight-speed Full-Text Retrieval using Character Transition Probability.
＊ Sueharu MIYAHARA, Fumihiko OBASI
＊ NTT Human Interface Laboratories

３．１ 索引の作成：　索引は文字単位にどの文にその文字が出現したかを記憶する方法(イ)、テキスト中の２文字の連接を索引にする方法(ロ)、テキストを言語解析して単語を抽出し、その単語から連接を調べて索引にする方法(ハ)を採用した。なお、平仮名は索引から除外し、その対策として平仮名以外の索引から得られた候補文に対して平仮名も含めてサーチする方法を採った。これら索引付けの方法とその時の索引の状態を表１に示す。

３．２ 検索方式と検索結果：　検索方式と検索時間との関係を表２に示す。これは検索文５例の平均を表わしたもので、１つの検索文あたりキー単語は平均５個、検索単語は平均１２単語である。方式Ａの検索では検索単語内に索引が複数個連続する場合、共通して出現する文番号を候補文と見なしてサーチ処理を行い高速化した。　測定の結果、全文をサーチして検索を行った場合に約１９.５秒を要したテキストが、言語解析した単語を用いて２文字の連接索引を作成した場合、約７倍の速度で検索でき、上位１０個の検索結果に変化がなっかた。これは検索文あたり候補文が平均５５０個と多かったためと考える。なおこのとき、テキスト容量の約８０％に相当するメモリが索引用として必要になった。

３．３ 考察：　① 索引をより細かな単位にすると、候補文が増加してテキストサーチの時間は増加するが、検索漏れを少なくできる利点がある。しかし、（イ）や（ロ）の方法では索引のデータ量の増加から平仮名を索引に持つことは困難と思われるが、（ハ）の方法は可能と考える。
② 上記以外にもテキストを言語解析して単語を索引にする方法が在るが、この方法は検索速度がやや速くなるが、部分一致の文字列を検索でき出来ないと云う問題が残る。

４．むすび　　フルテキスト検索を高速に行うため、検索対象テキストを言語解析して得られた単語から連接索引を作成して検索する方法を検討した。その結果、テキスト容量の約８０％のメモリが索引用として必要になるが、検索精度をほとんど低下させることなく、検索速度を全文サーチに較べ、約７倍に向上できることが分かった。　今後は索引作成の高速化やシステム操作や検索・確認の容易化について検討していく。

謝辞　　本研究の機会を与えて戴いた当研究所・川嶋言語ﾒﾃﾞｨｱ研究部長、協力戴いた加納英文氏に深謝する。

表２．検索方式と処理時間の関係

方式	処理（検索と順位付け）	索引	処理時間*（秒）
A	文字の指示する文をサーチ	イ	３.０６
B	連接の指示する文をサーチ	ロ	２.９４
C	連接の指示する文をサーチ	ハ	２.７３
D	全文サーチ	無	１９.５０

＊ 検索文５例の平均、　＊ 被検索文の長さは平均５５文字

文献 1)宮原他“文書情報蓄積検索…”, 平1情処全大,2N-6 (1989-10).
2)加藤他“大規模文書情報ｼｽﾃﾑ用ﾃｷｽﾄ…”, 情報学基礎,14-6 (1989-7).
3)日比野他“LISPﾏｼﾝELISの基本…”,情処記号処理研資,12-15 (1980).