7H－6

高速先頭照合方式による ストリングサーチ高速化の検討

秋沢　充　　野口孝樹　　川口久光
加藤寛次　　畠山　敦　　志村隆則
（株）日立製作所　　中央研究所

1．はじめに

情報処理システムの記憶装置の容量が増大するに従い，文書データを漏れなく高速に検索する処理が一層重要となっている。しかも，一般ユーザが簡単に文書データの蓄積･検索をしたいという要求がある。このため，自由に設定したキーワードによる高速の全文検索の必要性が高まっている[1][2][3]。

全文検索において重要なストリングサーチ･アルゴリズムのひとつとして，複数キーワードを一括して検索する有限状態オートマトンを用いた方法[4][5]（以下ＦＳＡ法と記す）が知られている。本報告ではＦＳＡ法をベースとしたストリングサーチの高速化方式について述べる。

2．ＦＳＡ法高速化の課題

図１にＦＳＡ法の処理の概念図を示す。一般に状態遷移テーブルはメモリで構成され，その容量は大きく，状態遷移を制御する部分とは別チップ構成となる。文書データが１文字入力されるたびに状態遷移テーブルをアクセスし，状態遷移のシーケンスが繰り返される。したがって状態遷移テーブルへのアクセスが毎サイクル必要となり，処理速度向上の妨げとなる。

今回我々は，状態遷移テーブルのアクセス頻度を減らすことで高速化が可能であることに着目し，新たな高速ストリングサーチ方式を検討した。

3．高速先頭照合方式の提案

本報告で提案する高速先頭照合方式は，各状態への遷移頻度の差に着目した方式である。ハードウェア論理によって遷移頻度の高い状態におけるテーブルアクセスを不要とする。これによりストリングサーチを高速化する。

図２にＦＳＡ法の解析モデルを示す。簡単のためにキーワードが“$C_0C_1\cdots C_k\cdots C_{n-1}$”（$C_k$：任意の文字）のみの場合について考察する。レベルｋにおいて状態遷移処理が行なわれる場合には，必ずレベル０，…レベル（ｋ－１）での処理が行なわれている。したがって，従来のＦＳＡ法ではレベル０が状態遷移テーブルのアクセス頻度が最も高く，順次レベル１，レベル２，…と低くなる。高速先頭照合方式ではレベルｋ＝ｋｃまでをハードウェアによって高速に処理（先頭照合処理）し，レベル（ｋｃ＋１）以降は有限状態オートマトンの状態遷移により処理（後方照合処理）するものとする。

［定義１］任意の状態での状態遷移処理において，これをソフトウェアで実行した場合のステップ数に換算したものを処理量と定義する。

［定義２］レベル０からレベルｋまでの処理量の総和の，レベル０からレベル（ｎ－１）までの全処理量に対する割合ＷＲ（ｋ）を相対部分処理量と定義する。

ある特定の文字（字種数Ｎｃ）が文書データ中に出現する確率を１／Ｎｃであるとすると，これは状態遷移確率ｆに等しいと考えることができる。このとき，相対部分処理量ＷＲ（ｋ）は次式となる。

$$WR(K)=\frac{1-f^{k+1}}{1-f^{n}} \qquad (1)$$

（$0\leq k\leq n-1$，ただし $f=1/Nc$）

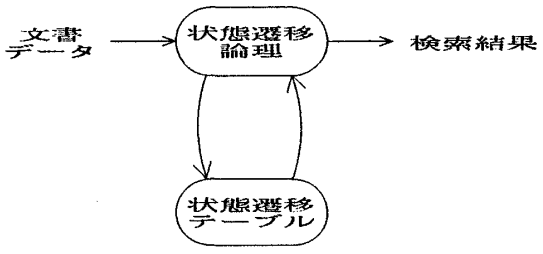

図１　ＦＳＡ法の処理概念図

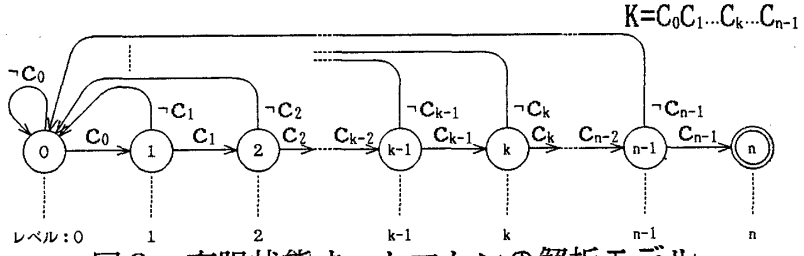

図２　有限状態オートマトンの解析モデル

Proposal of New String Search Method with Fast Partial Pattern Matching
Mitsuru Akizawa, Kouki Noguchi, Hisamitsu Kawaguchi,
Kanji Kato, Atsusi Hatakeyama, Takanori Shimura
HITACHI, Ltd.

“ａｂｃｄｅ”を検索する場合のＷＲ(ｋ)を，一例として求めた結果を図３に示す。ｋｃ＝２とすると全処理量の９９％を，ｋｃ＝１としても全処理量の９６％を先頭照合で処理できる。以上の評価から，キーワードの先頭の２～３文字の比較処理が全処理量の大部分を占めることがわかる。したがって，この部分をハードウェア化して高速に先頭照合処理をすることで，ストリングサーチが高速化される。

図４はｋｃ＝１の場合に，本方式を並列比較器と後方照合用に生成した有限状態オートマトンとで構成した概念図である。長方形で囲まれた２文字を並列比較器へ設定する。並列比較器からの一致信号がオートマトンの初期状態を発火させ，以後は状態遷移テーブルに従って文書データとの比較照合処理を行なっていく。

４．予測性能の評価

高速先頭照合処理方式を採用したストリングサーチの加速率の評価を行なう。ストリングサーチ処理全体の加速率Ａｔは次式より求められる。

$$1/At=(Rc/Af)+(1-Rc) \qquad (2)$$

ただし，$Rc \equiv WR(kc)$ (3)

ソフトウェア処理を基準とした，並列比較器による先頭照合処理の加速率Ａｆを８０，日本語文書における相対先頭照合処理量Ｒｃを９６％(ｋｃ＝１：先頭２文字)と仮定する。これらの仮定から相対ストリングサーチ処理時間(１／Ａｔ)を算出した。

算出結果から，日本語文書のストリングサーチにおいて，高速先頭照合方式(ｋｃ＝１：先頭２文字)により約２０倍の加速率が達成可能であることがわかった。

５．おわりに

ＦＳＡ法をベースとしたストリングサーチを高速化する高速先頭照合方式を提案した。従来のソフトウェア処理に比較して，約２０倍の加速率となる見通しを得た。

参考文献

[1]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)

[2]高橋,他:フルテキストサーチのハードウェア技術について,同上,pp.14.5.1-14.5.8(1989.7)

[3]早川,他:ストリームデータプロセッサSDP(1)(2),情処第37回全大,pp.113-116,(1988)

[4]A.V.Aho,M.J.Corasick:Efficient String Matching, Commun.ACM,Vol.18,No.6,pp.333-340 (1975.6),

[5]篠原,有川:日本語テキスト用のAho-Corasick型パターン照合アルゴリズム,情処研報,Vol.86,No.48,pp.52.4.1-52.4.8(1985.11)

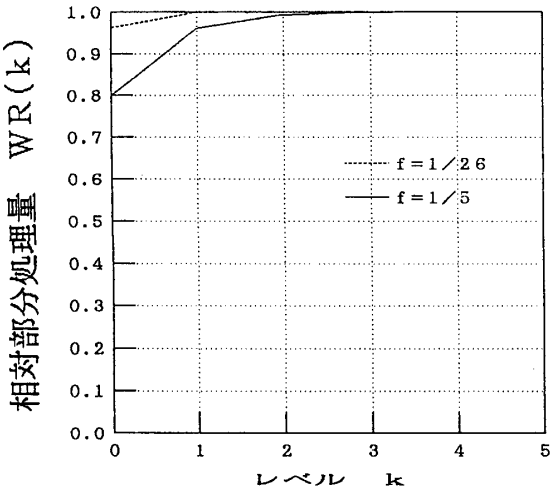

図３ 解析モデルによる
相対部分処理量

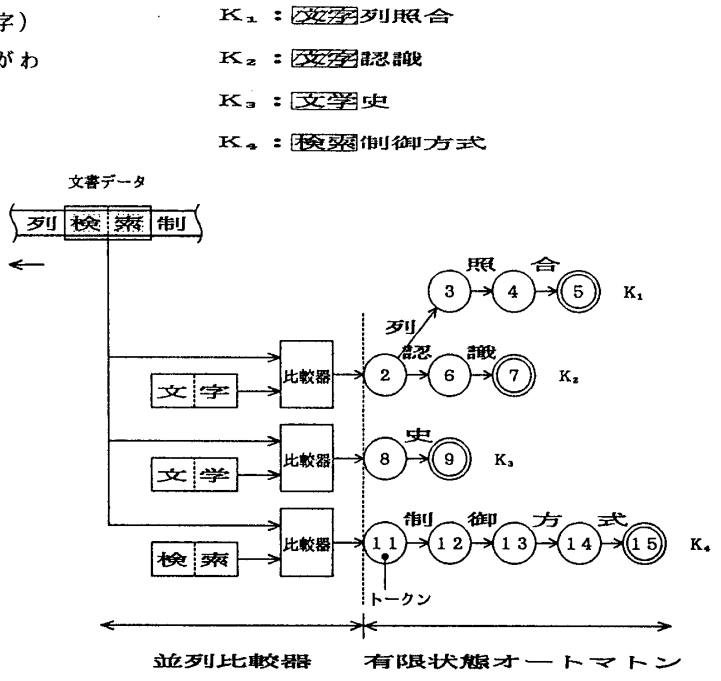

図４ 並列比較器とＦＳＡで構成した
高速先頭照合方式の概念図