1Q−01

BM法を用いたハードウェア方式による複数文字列を同時検索するアルゴリズムの提案

長田　千加子、大曽根　匡
専修大学 経営学部 情報管理学科

1．はじめに

【1】では、BM 法をハードウェアを用いて並列化した文字列検索アルゴリズムを提案した。本論文では、それを複数パターンの同時検索を可能にするように拡張する。

2．ハードウェア構成

単数パターンに対する検索から複数パターンに対する検索に拡張するにあたって、ハードウェア構成の変更した点を以下に示す。ここで、説明を容易にするため、パターン数 n は２とする。

①パターンレジスタの個数を２個にする（P_1 と P_2）。

②データレジスタ D とパターンレジスタ P_1 と P_2 の文字同士を同時比較できるように、並列比較ユニット PCU も２個にする。

③比較結果ラッチも C_1 と C_2 の２個用意し、比較結果を別々に記憶する。

3．動作例

例として、パターン１が「ABCD」、パターン２が「EFGH」の場合について考える。また、テキストは図２に示す。

この例の場合、あらかじめパターンレジスタ P_1 に「ABCD」、P_2 に「EFGH」を格納しておく。初期縦読みアドレスは３に設定しておく。

(1) 時刻１：縦読みアドレス３から最初の縦読みデータ「IMPS」をフェッチし、データレジスタ D に取り込む。そして並列比較ユニット PCU_1 と PCU_2 を用いてデータレジスタ D とパターンレジスタ P_1、P_2 の各文字の比較を行う。この場合、両者の文字同士で一致する文字はないので、比較ラッチ C_1 と C_2 は全て０となり、横読みを行わないことを判断する。そして、縦読みアドレスを＋１６更新し１９とする。

(2) 時刻２：縦読みアドレス１９から２回目の縦読みを行い、データ「JCNF」をデータレジスタ D に取り込む。このとき、データレジスタ D

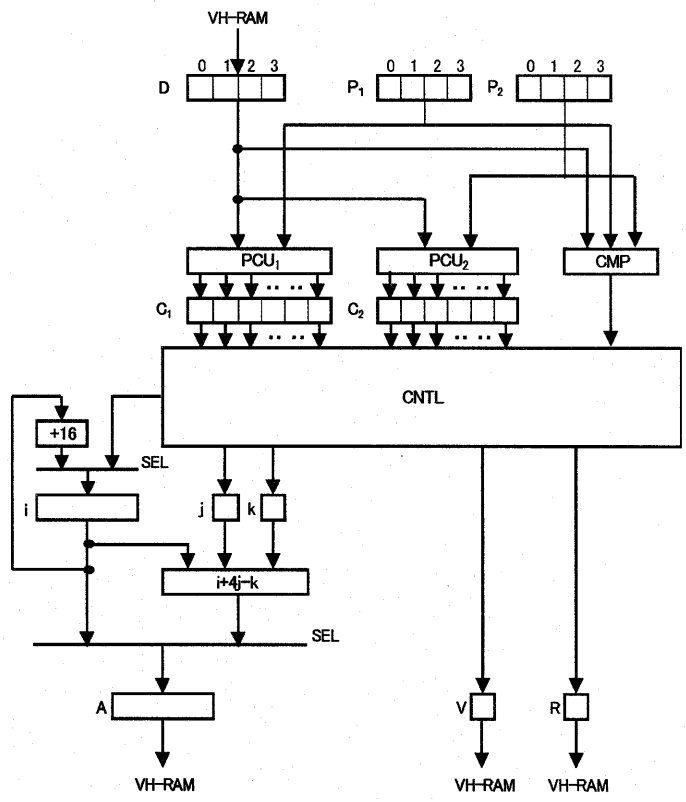

図１　検索エンジンのハードウェア構成

A Parallel String Searching Algorithm for Multiple Patterns Using BM Method.
Chikako Nagata and Tadashi Osone
Department of Information Management, Senshu University

の１文字目「C」とパターンレジスタ P_1 の２文字目の「C」が一致するので、ラッチ $C_1[1,2]$が１となる。そこで、横読みアドレス２１を算出する。同様に、ラッチ $C_2[3,1]$も１となる。これより、横読みアドレスは３０となる。

テキスト

0	A	B	C	I
4	F	A	B	M
8	C	D	A	P
12	B	C	D	S
16	C	A	B	J
20	D	A	B	C
24	D	D	A	N
28	A	B	E	F
32	G	H	A	V
36	B	C	D	L
40	U	R	B	B
44	D	B	K	Q

図２　テキストの例

（3）時刻３：横読みアドレス２１からテキスト「ABCD」を横読みし、データレジスタ D に格納する。この場合、パターンレジスタ P_1 とデータレジスタ D の内容が一致するのでパターン１を検出する。

（4）時刻４：同様に、横読みアドレス３０からテキスト「EFGH」を横読みし、データレジスタ D に取り込む。この場合もパターンレジスタ P_2 とデータレジスタ D の内容が一致しているので、パターン２を検出する。そして、すべての横読みが終わったので、縦読みアドレスを３５に更新し、縦読みを再開する。

（5）時刻５：縦読みアドレス３５からデータ「VLBQ」をデータレジスタ D に取り込む。このとき、ラッチ $C_1[2,1]$が１となる。そこで、パターンレジスタ P_1 に対して、横読みアドレス４２を算出する。

（6）時刻６：横読みアドレス４２からテキスト「ＢＢＤＢ」を横読みし、データレジスタ D に格納する。そして、パターンレジスタ P_1 と比較する。この場合は一致しないので、パターン１がここに存在しないことがわかる。

このようにして、最終的には、縦読み３回、横読み３回の計６回のデータ読み出しで、２つのパターンを同時に検索することができる。

４．性能評価

パターン数 n を複数にしたことによって、読み出し回数がどのように変化するかをシミュレーション実験した。その結果を図３に示す。縦軸に読み出し回数を、横軸にアルファベットの文字の種類数を示している。

アルファベットの文字の種類数は、４、８、１６、３２、６４、１２８と変化させた。テキスト長は、１００００文字とし、ランダムに発生させた。また、パターンについては、長さを４文字とし、ランダムに生成した。パターン数 n は、１から４まで変化させた。

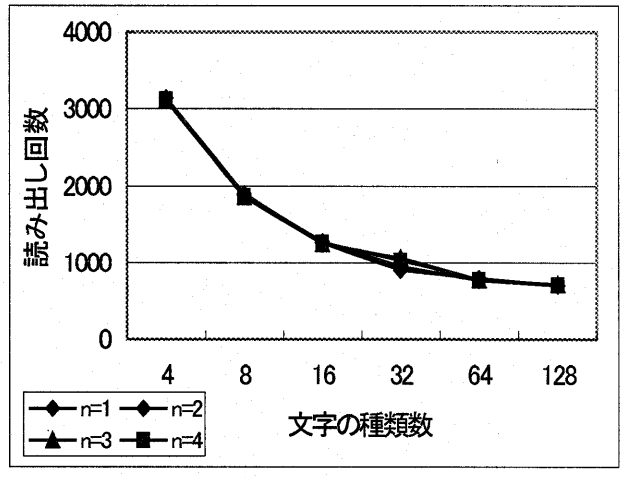

図３　読み出し回数の変化

この図から、文字種類数が増加することによって、読み出し回数が減少していることがわかる。そのことより、文字の種類数が増加するほど、アルゴリズムの性能が良くなることがわかる。

また、パターン数 n が増加しても、読み出し回数は大きくは変化しないことがわかった。

参考文献

【1】大曽根匡：「BM 法を用いたハードウェア方式による文字列検索アルゴリズムの提案」、IPSJ 第 61 回全国大会　1Q-02、2000.10.