1R−11　入力文字情報を最大限に活用した文字列検索アルゴリズムの提案

大曽根　匡、佐藤　創
(専修大学経営学部情報管理学科)

1．はじめに

近年、文献情報等の文書情報のＤＢ化が急速に進められているのに伴い、文書情報処理の高速化のニーズが極めて高まっている。このような状況の中で、文書情報処理のうち最も基本的かつ高負荷な処理の一つであるストリング・サーチ処理の高速化は重要な課題である。その高速化を目的としたアルゴリズムとして、ＫＭＰ法やＡＣ法、ＢＭ法などが著名である。また、ＡＣ法とＢＭ法をハイブリッドしたアルゴリズムもいくつか提案されている[1][2]。しかし、これらのアルゴリズムは、入力した文字の情報を十分に活用しきっているとはいえないものであった。そこで、本稿では、入力文字情報を最大限に活用した文字列検索アルゴリズムを提案する。また、性能実験により、その有効性についても検証する。

2．提案アルゴリズム

高速なアルゴリズムとして著名なＢＭ法は、パターンの末尾の文字から照合を始めることにより、高速化を図っている。例えば、パターンが「ＡＢＣＤ」の場合は、「Ｄ」から照合を始める。そのために「Ｄ」に対応する文字をテキストから入力する。もしこの入力文字が「Ａ」だったらパターンを右へ３文字分シフトし、再び、シフトしたパターンの末尾の文字に対応する文字を入力する。しかし、ＢＭ法では、パターンを右へシフトした際、それまでに入力した文字の情報を全て忘れてしまうようになっている。

そこで、提案するアルゴリズムでは、パターンを右へシフトした際にもそれまで入力した文字情報を覚えておくようにした。前述の例では、現在「Ａ＿＿＊」という照合状況であるとして覚えておくようにする。ここで、＿は未照合文字、＊は次の照合文字を表す。すなわち、この照合状況は、パターンの先頭文字の照合は済んでおり、つぎに４文字目の文字の照合を行うことを示している。

提案アルゴリズムでは、パターンから考えられる全ての照合状況をまず生成する。パターンが「ＡＢＣＤ」の場合の全ての照合状況を表１に示す。次に、[1][2]のアルゴリズムと同様に、状態ｉのときに文字ｃが入力されたら、次の状態が何になるかという情報が書かれた状態遷移テーブル T[i,c] と、次に何文字先の文字をテキストから入力すればよいかという情報の書かれたスキップテーブル S[i,c] を作成する。表２と３に、パターンが「ＡＢＣＤ」の場合の状態遷移テーブルとスキップテーブルを示す。そして、パターンの検索は、この２つのテーブルの参照を繰り返すことにより行われる。その動作例を図１に示す。初期状態は０とし、テキストの４文字目から文字を入力し始める。この例の場合、提案法では、25 文字のテキストに対し 10 文字の入力で検索を終了している。すなわち、平均スキップ幅は 2.5 文字である。一方、ＢＭ法では 14 文字の入力を要する。一般に、提案法は、原理的にＢＭ法やＫＭＰの性能を下回らないことに注意されたい。

3．性能実験

アルゴリズムの性能は、平均スキップ幅によって表現できる。そこで、提案法とＫＭＰ法、ＢＭ法の平均スキップ幅を定量的に比較するために、各種の実験を行った。そのうちの顕著な差の認められた結果を以下に示す。図２は、パターンを「ＡＢＣＤ」とした場合、テキスト上の「Ｃ」の出現確率に対し平均スキップ幅がどのように変化するかを示した図である。実験では、アルファベットの文字種は 32 文字とし、長さ 10000 文字のテキストをランダムに生成し、そのテキストに対する平均スキップ幅を求めた。ここで、「Ｃ」以外の文字の出現確率は一様とした。この試行を 100 回行い、その平均を求めた。これより、「Ｃ」の出現確率の増加に伴い、提案法の性能とＢＭ法の性能との差がだんだん大きくなっていくことがわかった。しかし、「Ｃ」以外の文字の出現確率が増加に対しては、ＢＭ法と提案法の性能にほとんど差が出ないという結果が得られた。図３は、パターンを「ＡＡＡＡ」とした場合の、テキスト上の「Ａ」の出現確率の変化に対する平均スキップ幅である。これより、「Ａ」の出現確率の増加に伴い、ＢＭ法の性能が著しく劣化するのに対し、提案法の性能はＫＭＰ法の性能を下回らないことが確認された。図４は、パターンを「１０１０１」とした場合の、テキスト上の「１」の出現確率の変化に対する平均スキップ幅である。ここで、アルファベットの文字種は「１」と「０」の２文字とする。このように、特にアルファベットの文字種の少ない場合に、提案法が有効であることがわかる。

A String Searching Algorithm with full use of Information obtained in Text Scanning

Tadashi OHSONE and Hajime SATO

Department of Information Management, Senshu University

表１．状態の定義

状態	照合状況
0	＿＿＿＊＿＿＿＿
1	＿＿＊D＿＿＿＿
2	＿＊CD＿＿＿＿
3	＊BCD＿＿＿＿
4	ABCD＿＿＿＊
5	＿＿C＊＿＿＿＿
6	＿B＿＊＿＿＿＿
7	A＿＿＊＿＿＿＿
8	＿B＊D＿＿＿＿
9	A＿＊D＿＿＿＿
10	A＊CD＿＿＿＿

＊：入力位置

表２．状態遷移テーブル

	A	B	C	D	#
0	7	6	5	1	
1			2		
2		3			
3	4				
4	7	6	5	1	
5	7			2	
6	7	6		8	
7	7	6	5	9	
8			3		
9			10		
10		4			

空白は状態０

表３．スキップテーブル

	A	B	C	D	#
0	3	2	1	-1	4
1	5	5	-1	5	5
2	6	-1	6	6	6
3	7	7	7	7	7
4	3	2	1	-1	4
5	3	4	4	-2	4
6	3	2	4	-1	4
7	3	2	1	-1	4
8	5	5	-2	5	5
9	5	5	-1	5	5
10	6	6	6	6	6

＃：その他の文字

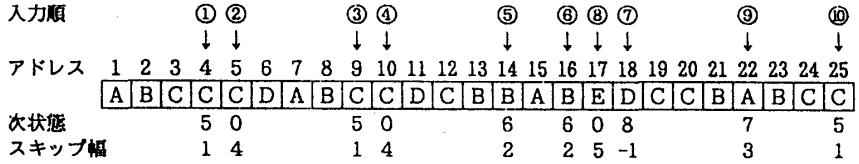

図１．提案法の動作例

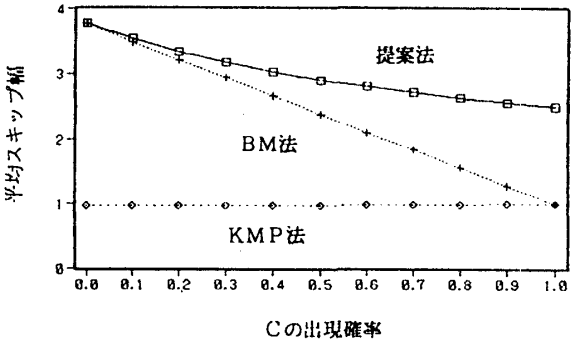

図２．パターン「ＡＢＣＤ」に対する平均スキップ幅

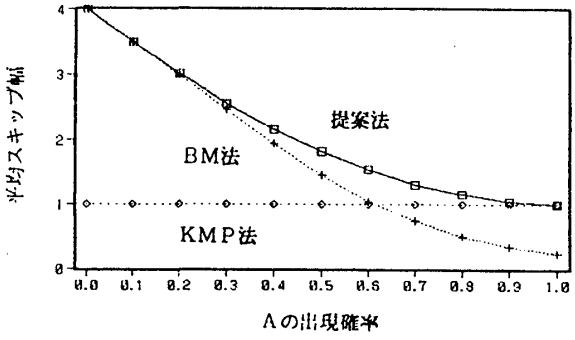

図３．パターン「ＡＡＡＡ」に対する平均スキップ幅

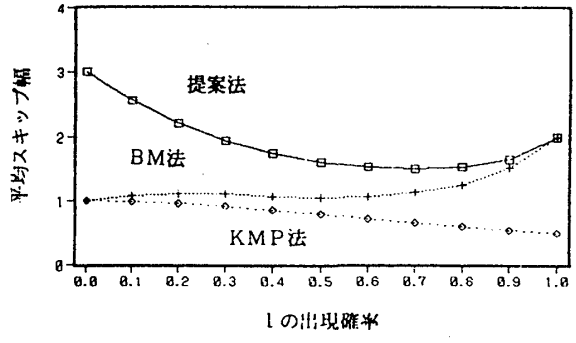

図４．パターン「１０１０１」に対する平均スキップ幅

４．まとめ

入力した文字情報を最大限に活用した文字列検索アルゴリズムを提案した。また、性能実験によりＫＭＰ法やＢＭ法より有効であることを定量的に検証した。特に、２種類のアルファベットからなるテキストにおける文字列検索に有効であると考えられる。

参考文献

【1】大曽根 他,"高速ストリング･サーチ･アルゴリズムの提案," 情報処理学会第34回全国大会論文集, pp.463-464 (1987).

【2】大曽根,佐藤,"文字の出現頻度を考慮した文字列検索アルゴリズムの提案," 情報処理学会第45回全国大会論文集(1), pp.67-68 (1992).