文字列探索アルゴリズム評価法の提案

1Q-03

大段雅典　佐藤匡正

島根大学大学院総合理工学研究科

1．序論

文字列の探索は，被探索文字列中に一致する探索文字列の有無やその位置を求めるものである．この処理では，一致比較が頻繁に行われるため，速度の向上を狙った様々なアルゴリズムが知られている．力まかせ(BF)法は最も単純な探索アルゴリズムであり，クヌース-モーリス-プラット(KMP)法はそれを改良したものである．さらに，KMP 法の逆走査版(BSKMP 法)及び，不一致文字(MM)法を組み合わせたボイヤ-ムーア(BM)法，ハッシュ法を用いるがハッシュ表もたないラビン-カープ(RK)法などが代表的なアルゴリズムである．

これらのアルゴリズムの評価では，探索時の文字の比較回数や計算量を基準とするのが一般的である．探索文字列長をm，被探索文字列長をnとすると，BF 法は約 nm 回，KMP 法は 2(n+m)回以下，BM 法は平均的な比較回数が n/m 回，RK 法はほとんど確実に n+m ステップである[1] とされている．しかしこの様な評価は統一された観点に欠ける上に，様々な種類の入力(文字列)に配慮されていないなどの問題点がある．そこで，本論文では，文字列探索アルゴリズムの評価法を提案し，実際にそれによる評価結果について論ずる．

2．評価

2.1　方法

代表的なアルゴリズムは 5 種類に整理できる．BM 法は，MM 法と BSKMP 法による複合アルゴリズムと考えることができる．

上に示したアルゴリズムに基づき，同一仕様の文字列探索プログラムを作成し，文字ごとの比較回数の総計を比較する．

さて，探索の速度は，基本的には与えられる入力(文字列)によるが，主として次のように整理できる．

① 入力列の種類
② 探索文字列と被探索文字列の長さ
③ 部分一致する文字数
④ 完全一致を検出する回数

入力列の種類は，ビット列，数字列，英文，和文の 4 種類とし，表 1 のように入力列種ごとに被探索列を固定する．ビット列と数字列については，任意に並べた 10 文字の数列を繰り返し用いて最大 1000 文字とする．英文と和文は表中の書籍から自然な文章を抜き出して 1000 文字とする．

表 1　文字列種別の被探索文字列

入力列種	被探索文字列
ビット列	(1111010011)*
数字列	(2098215045)*
英文	Object Oriented Programming
和文	帝都物語　序

ビット列と数字列は，部分一致の割合やその開始点が異なるように数種の探索列を作成する．英文と和文は，被探索列中に存在する適当な語句を探索列とする．ビット列と数字列は 20 種類ずつ，和文と英文は 10 種類ずつとする．文字はアルゴリズムの特性上，3 個以上とする．以下に数字列の探索文字列の一部を示す．

表 2　数字列の探索文字列

	探索文字列
一致文字無し	333
0 文字目から 2 文字一致	208
6 文字目から 3 文字一致	15044
0 文字目から 9 文字一致	2098215044
10 文字目から前に 2 文字一致	945
5 文字目から前に 2 文字一致	19882
10 文字目から前に 9 文字一致	1098215045

2.2 比較回数

探索における比較回数は次で与えられる.

(1) RK 法を除くアルゴリズム

比較開始文字からの一致文字数を M_i,

探索文字列の移動回数を P とすると,

$$比較回数 C = \sum_{i=1}^{P}(M_i + Q)$$

ここで, 完全一致のとき Q=0, 不一致のとき Q=1

(2) RK 法

$$比較回数 C = n - m + 1 - S*(m-1)$$

ここで, S は完全一致検出回数.

(3) 実測

4 種類の被探索文字列に対し様々な条件の探索文字列を探索した結果を図 1 に示す. これは, ビット列で1字目から3文字一致するという{11100}を探索し比較回数を計測した結果である. これから, 被探索列長ごとのアルゴリズムの比較回数がわかる.

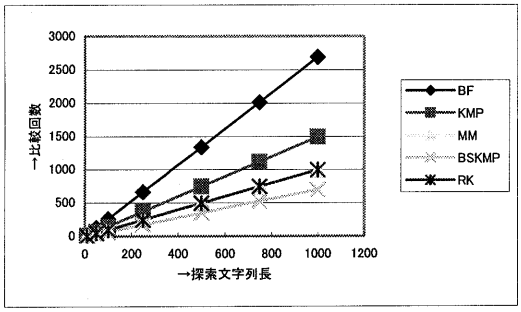

図 1 ビット列{11100}探索時の比較回数

2.3 分析

(1) 図 2 に, ビット列における BF 法に対する比較回数比をアルゴリズムごとに求めた結果を示す. これから, 他のアルゴリズムが必ずしも BF 法より有効とは言えない.

(2) MM 法はもっとも低速とされる BF 法よりも比較回数が多い場合がある. この理由は次である.

① 逆走査法は, BF 法とは違い, 照合する走査が逆である. 従って, 同じ探索列でも, 一度の探索で部分一致する文字数が, BF 法で探索する場合よりも多いことがある.

② 入力列種がビット列では, 文字が{0, 1}の 2 種類しかないために, 探索列の移動文字数が多いという特性が生かされない.

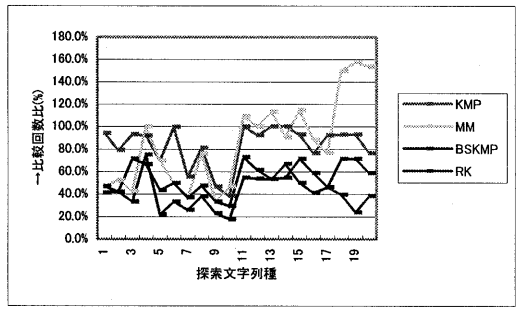

図 2 ビット列における比較回数比(対 BF 法)

(2) 図 3 は, 全ての入力列における比較回数比を示している. これは, それぞれのアルゴリズムの比較回数が, BF 法を基準としたときに, 0. 9 以下である場合の割合を求めて, 入力列種ごとに積み上げている.

このとき, 英文と和文の自然言語での文字列探索において, アルゴリズムの差が顕在化している. 文字種の増加に伴い部分一致部分が少なくなるため, 必ずしも KMP 法が有効とは言えない. それに対し, MM 法については英文, 和文の文字列探索には有効であると言える. 一方, BSKMP 法はビット列と数字列に有効であるから, 比較回数は最小である.

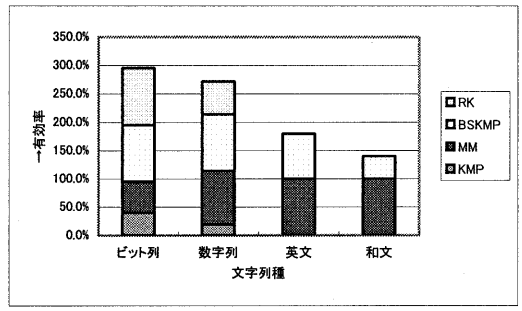

図 3 アルゴリズム別の有効度

文字比較回数による本評価法では, RK 法を対象としない. これは, RK 法は文字種に依存しないため, 比較回数が一定になるからである。

3. 結論

現実的な評価方法を提案し, その評価結果からビット列, 数字列の探索では BSKMP 法, 英文, 和文の文字列探索では MM 法が最も有効であることが言える.

参考文献

1) Robert sedgewick "Algorithm in C", Addison-Wesley Publishing Company, Inc. (1990).