研究課題別評価

1 研究課題名：
数値演算的手法による文字列処理

2 研究者氏名： 篠原 歩
ホスドク研究員： 馬場 謙介 （研究期間 H.14.4～H.15.10）
ポスドク研究員： 稲永 俊介 （研究期間 H.15.4～H.15.9, H.16.10～H.16.12）
ホスドク研究員： Heikki Hyyro （研究期間 H.16.2～H.17.1）

3 研究の狙い：
インターネットの発展を背景として膨大なデータが蓄積されているが、特に XML フォーマットの導入は、あらゆる情報を計算機と人間の双方に可読な文字列として表現することによって協調的な知的生産性を高めようとするものであり、大量のテキストデータを効率よく処理するための技術はますます重要になっている。パターン照合をはじめとした文字列処理に関する研究は、これまでも深く研究されてきたが、そこで用いられる手法は文字列としての性質を巧みに利用して文字列処理に特化されたものであり、他方で膨大な研究の蓄積がなされている数値演算とはほとんど無関係である。テキストデータの分類や、大量のテキストデータに潜む規則の発見など、知的な文字列処理を行うための操作の大部分は計算量的に困難な問題を内包しており、これらを現実的な時間で近似していくためには、「文字列」に対する新しい視点が必要であると考える。本研究は、文字列として与えられる情報を、そのままの離散的な値としてではなく、連続的な数値としてとらえるための枠組みを開発し、その上でさまざまな文字列操作を数値演算として実装できるようにするための基礎理論の構築を目的としている。

4 研究成果：
(1) 高速フーリエ変換による近似文字列照合の高速化
文字列処理の最も基本的な操作であるパターン照合は、テキスト文字列 T とパターン文字列 P が与えられたきに、テキスト T に現れるパターン P の出現位置を全て見つける問題として定式化できる。パターンとテキストの間の少数の不一致を許した近似文字列照合問題は、実用的な観点から極めて重要であり、本質的には T の各位置から始まる部分文字列と P との一致具合を表すスコアベクトルを計算する問題に帰着できる。本研究では、スコアベクトルの計算をテキストとパターンに対応する2つの関数の畳み込みとみなし、デジタル信号処理の分野でよく用いられる高速フーリエ変換(FFT)を応用することによって計算時間を落とすことに成功した。文字の種類が 2 つだけのときは、単に文字列を{-1,1}上の数値列に変換することで本質的には解決できるが、文字種が3以上になると、数値としての演算と、文字比較としての演算にうまい対応関係が取れなくなってしまう。この困難を克服するために、本研究では確率アルゴリズムを提案し、そこで得られる推定値が求めるスコアベクトルに一致することを証明し、またその分散を厳密に解析することに成功した。このように、数値処理でよく用いられる FFT の応用によって、近似文字列照合を効率よく行うことが原理的に可能であることを明らかにした。
(2) 一変数の文字列方程式の解の上限
数値的な手法において、与えられた制約を等式や不等式の形で表現し、それを満たす解を探すというのは最も基本的な操作である。また、観測データをうまく説明する法則を微分方程式という形で見つけ出すこともよく行われている。これらと同様に、文字列処理の最も基本的な操作である、文字列照合問題を、文字列に関する方程式とみなすことができる。例えば、テキスト文字列 abababaabbababa の中に出現するパターン aab の位置を探す問題は、 XaabY = abababaabbababa という文字列方程式を満たす変数代入 (X:=abab, Y:=bababa)を見つける問題とみなせる。より一般

的には、X aabY = aYabbaba のように、両辺に変数を含むような方程式を解くことを考える（この場合は X:=ababa, Y:=baba という解がある）。このような文字列方程式を解く問題は、見かけほど簡単ではない。例えば方程式 YYXbaaaaZ = ZYYXaaaab の最小解は、

X := aaaabaaaa,
Y := aaaabaaaaaaaabaaaabaaaa,
Z := aaaabaaaaaaaabaaaabaaaaaaaabaaaaaaaab

であり、問題の長さに比べて極めて長い。一方、文字列方程式 XabYY = bZZYX は解なしであり、文字列方程式を解くアルゴリズムを開発する際には、どの長さまで調べればその問題には解がないと言えるのかを知る必要がある。このように、解が存在するときにその最小解の上限を求めることが、そのまま最悪時の計算時間の評価に直結する。しかし、一般の文字列方程式に対する最小解の上限について、まだ十分にその挙動が解析されていない。そこで本研究では、変数が1種類だけ使われているような文字列方程式（例えば XabXbaaaaaa=aaaaaabaXbX）に限定して、最小解の長さの上限を求めることにした。既存研究においては、方程式の長さの4倍程度でおさえられるだろうという予測はなされていたが、その厳密な値は知られていなかった。我々は、網羅的に1変数文字列方程式を枚挙しながらその解を求めるプログラムを作成して長時間動作させ、その結果を分類しながら問題の分析を行った。その過程の中で、文字列の繰り返し構造のもつ性質に着目することによって、最終的には解の上限を厳密に、しかも極めて単純な式によって表すことに成功し、数学的な証明を与えた。また逆に、任意に指定された長さに対して、この上限と一致する問題を構成的に与えるアルゴリズムも示した。これらの論理的帰結として、1変数文字列方程式は、解をもつとすれば、その最小解は高々、問題の長さで抑えられることになる。このことにより、実用的な観点から、1変数文字列方程式を解くアルゴリズムの高速化が行えることがわかった。

(3) 高速パターン発見アルゴリズム

数値的に得られた観測データから、そこに内在する関係を関数として抽出する手法は、補間や内挿、外挿として知られ、数値データ処理の基本として、さまざまな関数族に対するアルゴリズムが開発され、実用に供している。近年、HTML や XML のような半構造を持ったテキストデータや、DNA 配列、アミノ酸配列のようにほとんど構造の知られていないテキストデータが大量に蓄積され、利用可能になってきたことをうけて、これらのデータベースから、そこに内在する関係を文字列パターンとして発見するための手法の開発が望まれている。本研究では、この問題を、2つの文字列集合を与えられたときに、それを高精度で分離する文字列パターンを見つける最適化問題として定式化し、そのために有用なデータ構造とアルゴリズムを与えた。まず、探索アルゴリズムに関しては、部分文字列パターン、部分列パターン、変数を含むパターン等、さまざまなパターン族に対して、その形式言語としての特徴を生かした探索空間の枝刈り技法を取り入れることにより、ユーザの指定するスコア関数を最適化するパターンを実用的な時間で発見できるアルゴリズムの開発に成功した。また、この中で頻繁に用いられるパターン照合を高速に行うための索引構造として、有効無閉路文字列グラフ(DAWG)，有向無閉路部分列グラフ（DASG）等を対象として，その性質や高速な構築アルゴリズムを開発した。

5 自己評価：

文字列処理技術の進展という観点から、おおむね順調に成果が得られたと考えている。特に、最初のハードルであった高速フーリエ変換の適用方法を検討する過程で、文字の差異と数値の差異をどう対応づけるかという本質的な問題に直面したが、この解決策として打ち出した確率的な手法は、今後も他の問題に適用できそうな手応えを感じている。また、文字列方程式に関しても、1変数という非常に限定された状況下ではあるが、解の長さに関する厳密な上限を与えることができたことに満足している。なお、数値方程式においては、解が存在しない場合や、厳密解の計算が困難な場合には、近似解を求めたり、（2次方程式に対して複素数を考えるように）解空間を拡大したりすることによって解決を図る。これに習って、文字列方程式に対する「近似解」や、解空

間の拡大についての考察を重ねてきたが、この方向には今のところ成果に結びついていない。数値演算の活用による文字列処理の高速化として、ビット演算命令を駆使した処理を開発したが、この手法は近年のCPUの持つベクトル命令（MMXやSSEなど）の有効活用に繋がる。しかし浮動小数演算コプロセッサの活用については、今後の課題として残った。

また、本研究の「協調と制御」領域としての意義付けを模索する中で、2年目から複数ロボットの協調制御としてロボカップ・サッカーへの応用に新たに取り組んだ。この2年間で戦果としては大きく前進したものの、本研究で得られた理論的な成果がまだ応用には直接的には結びついていないのが現状である。数値列の処理と文字列の処理の相違を意識しながら、有効な技術を相互に活用する方向で、この研究も継続していきたい。このように、それぞれの方向に進展があった分、また新たな課題をかかえることになっているが、本研究に取り組む前と比較して、より高い視点から問題設定を行い、解決策を探ることができるようになったと実感している。

なお、ポスドク参加型として、他の同様な制度と比較しても、より有利な条件で雇用関係を結ぶことができたおかげで、極めて優秀な内外のポスドクを迎え入れ、研究を推進する大きな原動力となった。本制度に心より感謝する次第である。

6 研究総括の見解：

文字列照合は、遺伝子配列の検索やインターネット検索のように膨大なデータベースからの検索時に必要不可欠で重要技術であるが、その計算量を削減する方法としてこのことに数値的演算手法を用いることの可能性を追求した。その結果、1変数の文字列方程式の解の上限を得るなど基本的な問題を解明すると共に、実用的には重要性の高い近似文字列照合に高速フーリエ変換の手法を応用して検索時間を短縮する試みを行うなど、基礎数学と実用的アルゴリズムの距離を短縮して、極めて独創性の高い成果を生み出したことは高く評価される。

7 主な論文等：

[1] Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda, and Setsuo Arikawa, "Compact Directed Acyclic Word Graphs for a Sliding Window", Proc. 9th International Symposium on String Processing and Information Retrieval (SPIRE2002), Lecture Notes in Computer Science 2476, pp. 310-324, Springer-Verlag, September 2002.

[2] Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda, Hideo Bannai, and Setsuo Arikawa, "Space-Economical Construction of Index Structures for All Suffixes of a String", Proc. 27th International Symposium on Mathematical Foundation of Computer Science (MFCS2002), Lecture Notes in Computer Science 2420, pp. 341-352, Springer-Verlag, August 2002.

[3] Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, Masayuki Takeda, and Setsuo Arikawa, "Discovering Best Variable-Length-Don't-Care Patterns", Proc. 5th International Conference on Discovery Science (DS2002), Lecture Notes in Computer Science 2534, pp. 86-97, Springer-Verlag, November 2002.

[4] Kensuke Baba, Ayumi Shinohara, Masayuki Takeda, Shunsuke Inenaga, and Setsuo Arikawa, "A Note on Randomized Algorithm for String Matching with Mismatches", Nordic Journal of Computing, Vol. 10, pp. 2-10, 2003.

[5] Kensuke Baba, Satoshi Tsuruta, Ayumi Shinohara, and Masayuki Takeda, "On the Length of the Minimum Solution of Word Equations in One Variable", Proc. 28th International Symposium on Mathematical Foundations of Computer Science (MFCS2003), Lecture Notes in Computer Science 2747, pp. 189-197, Springer-Verlag, August 2003.

[6] Heikki Hyyro, Jun Takaba, Ayumi Shinohara, Masayuki Takeda, "On Bit-Parallel Processing of Multi-byte Strings", Proc. 1st Asia Information Retrieval Symposium (AIRS2004),

October, 2004.
査読付き論文26（上記含む）、口頭発表2