様式Ｆ－１９

# 科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　25 年　5 月　20 日現在

機関番号：１１３０１
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：23650002
研究課題名（和文）　文字列学のための研究支援システムの開発

研究課題名（英文）　Development of A Research Support System for Stringology

研究代表者
篠原　歩（SHINOHARA AYUMI）
東北大学・大学院情報科学研究科・教授
研究者番号：00226151

研究成果の概要（和文）：
文字列の持つ離散構造や組み合わせ的性質の解明とその活用を効果的に行うための研究支援システムの開発を目指して研究を展開した．まず，種々の文字列の基本的性質や代表的アルゴリズムと索引構造を記載したオンラインシステムのプロトタイプを作成した．また，文字の置換を許した検索をサポートするための索引構造を開発した．さらに，極大な繰り返し構造である連を最も多く含む文字列を効率よく探索するアルゴリズムを実装した．

研究成果の概要（英文）：We developed a research support system for Stringology, that deals with algorithms and data structures used for string processing. We designed and built a prototype of an online encyclopedia of strings, where we explained various fundamental properties,　and typical algorithms and data structures on strings. We also showed an indexing structures that is suitable for parameterized pattern matching. We implemented efficient algorithms to find strings containing many runs, which are maximal repetitions in strings.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 交付決定額 | 2,200,000 | 660,000 | 2,860,000 |

研究分野：総合領域
科研費の分科・細目：情報学・情報学基礎
キーワード：アルゴリズム，情報基礎，離散構造，文字列学，データ構造

## １．研究開始当初の背景

記号の列としての文字列は，基本的な構造であり，計算機に情報を格納するための最も基礎的な役割を担っている．インターネット上にアクセス可能な膨大な HTML ファイルはもとより，あらゆるデータをタグ付きの文字列として表す XML 化によって，効率のよい文字列処理の研究の重要性は益々高まっている．一方，整数，有理数，実数，複素数などの数体系もまた基本的な構造であるが，数体系に対しては，Mathematica や Maple に代表される完成度の高い数式処理システムによって，微積分や行列計算等，代数的・数値的な処理が効果的に行えるようになっている．そのため，研究者は等式の変形や方程式の求解などのルーチンワーク的な作業は安心して計算機にまかせることができ，対話的な操作を通じて解くべき問題に専念することができる．また，研究を展開する中で出現した数列の性質やその一般項を知りたいときには，AT&T 研究所の N.J.A. Sloane が提供する「数列の事典」によって，既存の数列を容易に参照することができる．文字列に対しても，数体系や数列と同様な，研究支援ツールとしての文字列処理システムや，容易に参照できる「文字列の事典」が必要であるとの着想に至った．これまでは必要に応じて随時作成してきたプログラムを整理・統合し

てライブラリ化し，文字列の性質を容易に調べられる計算機環境と，既存の文字列の有用な性質を体系的にまとめたオンラインのデータベースを開発することが，次世代の文字列処理研究には必須であると考え，本課題に着手した．

２．研究の目的

本研究は，文字列の持つ離散構造，組み合わせ的性質の解明とその活用を効果的に行うための研究支援システムの開発を目指すものである．このシステムの構築に必要な理論的基礎を体系的に整理するとともに，プロトタイプを作成しながらその方向性を見極めることを具体的な目標とする．本研究は，文字列の組み合わせ的性質やその活用に関する学術的な研究を支援する計算機ツールを構築するために，その組み合わせ的性質そのものを活用する点にその特色がある．例えば，文字列の代数的性質はそれを構成するアルファベット（例えばΣ＝{0,1}やΣ’＝{a,b}）には依らないが，既知の文字列と比較する際には，その対応を取りながら照合する必要がある．また，長い文字列の性質を調べる際には，索引構造の構築や圧縮など，効率的な処理を行わなければ対話的な操作感が得られない．本研究では，それぞれの操作において，効率の良さを重視しながら実装を進めていくことで，その有用性の検証も兼ねる．そして，これらのツールを利用しながら，文字列の基本的性質を解明していくこともまた具体的な目的の一つである．

３．研究の方法

本研究は，文字列の組み合わせ的性質や構造を研究するための支援ツールと「文字列の事典」のプロトタイプを作成してその有用性を検証することを目指すものであり，関連する種々の要素技術を含む，下記の３つの研究項目に取り組んだ．

(1) 文字列の処理システムの開発

数式の処理において極めて有用な数式処理システムである Mathematica や Maple にも，最も基本的な文字列照合程度の関数群は提供されているが，文字列学の研究を実際に展開していくためのツールとしては全く不十分である．特に，繰り返し構造の検出や，Lempel-Ziv 分解などは，高次の文字列処理に必須の操作であるが，既存のパッケージでは提供されていない．また，文字列の索引構造としての接尾辞木，接尾辞トライ，有向無閉路グラフなどについては，それを実装するだけでなく，木構造・グラフ構造として対話的に可視化してみせることが文字列の性質を理解する上で極めて重要である．そこで，これまでに我々が必要に応じて個々に開発してきたプログラム群を再構成し，統一的な操作感にまとめ，また Web アプリケーションとしてのユーザインターフェースを提供するためのプロトタイプを作成した．

(2)「文字列の事典」サーバの構築

「数列の事典(http://oeis.org)」は，入力フォームに例えば「2,3,5,8,13」という数列を入力すれば，この数列を部分列として含む有名な数列をすべて検索結果として返してくれる．この場合は，フィボナッチ数として知られる数列 (0,1,1,2,3,5,8,13,21,...) の一般項やその出典へのリンク，基本的性質等が表示される．一方，abaababaabaab... という文字列は，文字列学の分野ではフィボナッチ文字列としてフィボナッチ数と関連した種々の興味深い性質を持つことが知られているが，このような代表的な文字列群を検索するシステムは存在しなかった．そこで，「文字列の事典」のプロトタイプを作成し，その基本的な仕様の策定や有用性を確認した．

(3)文字列の組み合わせ的性質の解明

上記システムの構築は，文字列学の研究に有用なツールの提供を目的としたものであるため，実際に我々が取り組んでいる種々の問題に適用することによってその効果を検証する必要がある．本研究では，文字列の連の数の上界と下界の解析をその具体的な対象として選んだ．文字列の連とは，その中に同じ文字列が２回以上続けて現れた極大な部分文字列をいう．例えば，aababaaa という文字列には，aa, ababa, aaaa という３つの連が含まれている．「長さ $n$ の文字列の中に，最大でどれだけ多くの連を含むことができるか」という基本的な文字列の組み合わせ的性質について，近年，大きな研究の進展があり，関心が高まっている．連の最多数が文字列の長さの線形で抑えられること，すなわち $0(n)$ であることが 1999 年に示されて以降，その係数である定数をより精確に求める研究が続けられており，現在の最良のものは上上界が 1.029,下界が 0.9445 である．また深く関連した連の指数和についても併せて解析を行う．例えば連 aa, ababa, aaaa の指数はそれぞれ 2, 2.5, 4 であるので，その指数和は 8.5 となる．この上界と下界をより精確に評価するために，効率のよいアルゴリズムを開発・実装し，計算機実験を行う．これらの作業を通じて，種々の文字列の組み合わせ的性質の解明と数値的評価を行った．またこうして得られた結果を文字列の事典に反映させた．


図 1　オンラインの文字列の事典

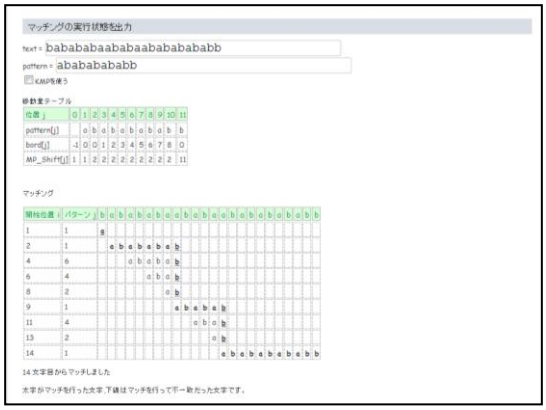


図 2　文字列照合の表示例

４．研究成果

(1) まず，システムの全体像についての検討を行いながら，これまで単発的に開発してきた種々のプログラム部品を整理・統合する作業を行った．スクリプト言語としては基本的には Python を用いることにし，高速な処理が必要な部分については C および C++ライブラリをリンクさせるという方針を決めた．また，データの電子化に関してもさまざまな現行技術を検討した結果，さまざまなフォーマットに柔軟に出力できることから，定評のある Sphinx システムを利用し，対話的な操作と図の描画のために HTML5 と JavaScript を用いることとした．このことにより，後述の通り，文字列学の研究活動にも有用な仕組みができた．

(2) オンラインの「文字列の事典」のプロトタイプを作成した（図１）．この中には，主に次の 3 種類のコンテンツを掲載している．①フィボナッチ文字列など，文字列学の対象となる種々の文字列についての定義や興味深い性質の説明．②パターン照合や周期の計算など，文字列処理の基本アルゴリズムの解説．③接尾辞木や接尾辞配列など，文字列処理に有用な索引構造の説明．

このシステムは，JavaScript を用いてユーザの入力に応じて動的に即座に表示を切り替える仕組みを持っており，上記①②③それぞれの項目で，この機能を活用することができる．基本アルゴリズムの解説である②において，例えばパターン照合については，図２のように，ユーザがテキスト文字列とパターン文字列を自由に入力することができる．この入力を替えると，瞬時に表示が切り替わり，内部で用いるテーブルの値やマッチングの様子を容易に確かめることができるようになっている．アルゴリズムによって照合の効率は大きく変化するが，ユーザはこれをさまざまな文字列に対して適用してみることで，その違いを体感することができる．
また，③のデータ構造の説明に関して，文字列処理において極めて多岐に渡って活用されている接尾辞木について，その性質や挙動を把握するためには，ある程度長い文字列に対して木を構築して観察する必要がある．しかしながら，この作業を手作業で行うのは大きな手間がかかった．そこでこのシステムには，接尾辞トライ（図３）や接尾辞木，有向無閉路文字列グラフ(図４)など，種々の有用な索引構造をリアルタイムに描画する機能を付けた．このことにより，データ構造そのものの理解を深めることに役立つのみならず，対象とする文字列の性質を視覚的に理解することにも有用であることが実感できた．また，後述のパラメータ化ポジションヒープという新たな索引構造の開発にも有用であった．今後の本格稼働には，さらなるコンテンツの充実が必要であるが，複数の作業者で協力しながらシステムを維持・更新できるように設計しており，また電子書籍のフォーマットへ変換することも容易である．

一方，数列の事典とは異なり，文字列の検索に関しては，文字の置換を許しながら照合を行う必要がある．例えば，xyxxyxyx と 10111010 は，字面上は異なるが，文字列としての組み合わせ的性質としては同等である．このように異なったアルファベット上で表現された文字列が入力されたとしても，同様な結果を返す必要がある．すなわち，通常の単なる文字列照合では不十分である．このような照合は，パラメータ化照合・置換不変照合など呼ばれ，近年大きな関心が寄せられている．本研究では，そのために有用な索引構造として，パラメータ化文字列に対応できるポジションヒープを新たに定義し，それを効率よく構築するアルゴリズムを開発することができた．

(3) 文字列の基本的な繰り返し構造である「連」を最も多く含む文字列を効率よく探索するアルゴリズムを実装した．このことにより，長さ 64 まで，連の最多数を厳密に求めることができた．さらに，指数和の下界を更

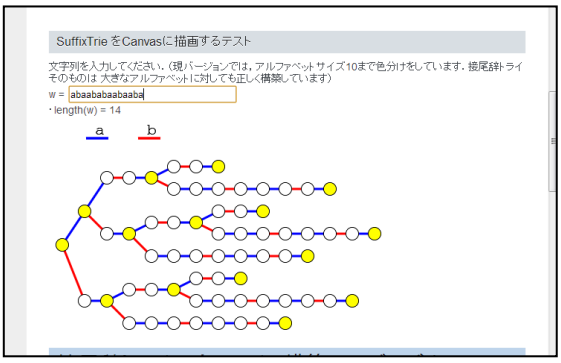


図 3 接尾辞トライ木の表示例

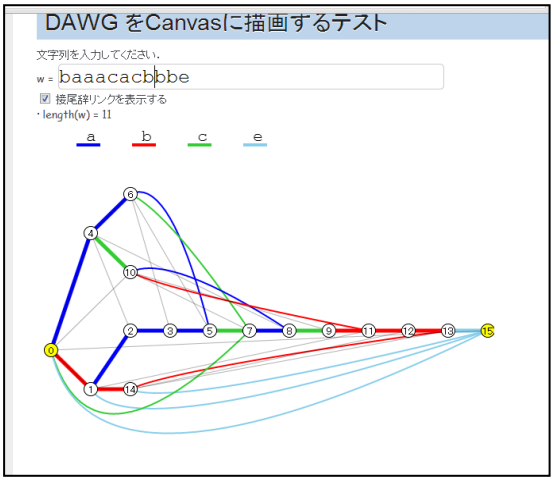


図 4　有向無閉路文字列グラフの表示例

新する新たな文字列の系列を発見した．この研究過程においても，この「文字列の事典」を対話的に用いて考察を容易に検証できることの有用性が検証できた．

以上に述べたとおり，システムのプロトタイプが作成でき，その効果が確かめられた．今後も，さらにコンテンツを充実させ，新たな機能を取り込みながら，システムを随時，更新していく予定である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

(1) Takashi Katsura, Kazuyuki Narisawa, Ayumi Shinohara, Hideo Bannai, Shunsuke Inenaga, “Permuted Pattern Matching on Multi-track Strings”, Proc. 39th International Conference on Current Trends in Theory and Practice of Computer Science, LNCS 7741, pp. 280-291, 2013, 査読有
DOI:10.1007/978-3-642-35843-2_25

(2) Kazuhiko Kusano, Kazuyuki Narisawa, Ayumi Shinohara, “Computing Maximum Number of Runs in Strings”, Proc. 19th International Symposium String Processing and Information Retrieval, LNCS 7608, pp. 318-329, 2012, 査読有
DOI:10.1007/978-3-642-34109-0_33

〔学会発表〕（計 6 件）

(1) 大田裕之，桂 敬史，成澤和志，篠原 歩，「マルチトラックデータ上の近似順列パターン照合と索引構造」電子情報通信学会コンピュテーション研究会，pp. 9-16, 2013 年 4 月 24 日，神戸大学

(2) 草野一彦，奥田遼介，成澤和志，篠原 歩，「文字列に含まれる連の最大指数和の解析～n=57 までの厳密値と新たな下界 2.03696 の発見」，電子情報通信学会コンピュテーション研究会，pp. 17-24, 2013 年 4 月 24 日，神戸大学

(3) 大友雄平，成澤和志，篠原 歩，「種々のパターン照合問題に対するポジションヒープの構築」，電子情報通信学会コンピュテーション研究会，2012 年 12 月 10 日，九州大学

(4) 相原高雄，篠原 歩，成澤和志，圧縮文字列に対する省メモリなパターンマッチアルゴリズム，電子情報通信学会コンピュテーション研究会，2012 年 10 月 31 日，東北大学

(5) 桂 敬史，成澤和志，篠原 歩，坂内英夫，稲永俊介，「マルチトラック文字列の順列パターン照合と索引構造」，電子情報通信学会コンピュテーション研究会，2012 年 9 月 3 日，法政大学

(6) 桂敬史，成澤和志，篠原歩，「マルチトラック文字列に対するパターン発見について」，夏の LA シンポジウム，2011 年 7 月 19 日，ザヴィラ浜名湖

〔その他〕

ホームページ等

http://www.shino.ecei.tohoku.ac.jp/stringology/

６．研究組織

(1)研究代表者

篠原　歩（SHINOHARA AYUMI）
東北大学・大学院情報科学研究科・教授
研究者番号：00226151