様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月２３日現在

機関番号：１７１０２
研究種目：若手研究（Ｂ）
研究期間：2009～2010
課題番号：21700019
研究課題名（和文）　パラメタ化文字列照合技法とパタン発見への応用
研究課題名（英文）　Parameterized string matching and its application to pattern discovery

研究代表者
稲永　俊介（INENAGA SHUNSUKE）
九州大学・システム情報科学研究院・特任准教授
研究者番号：60448404

研究成果の概要（和文）：パラメタ化文字列照合アルゴリズムは，自然言語テキストや生物学的配列上のパタン検索・発見を効率よく行うための基盤技術である．本課題では，パラメタ化文字列照合問題を高速に解くためのアルゴリズムの設計と，データ構造の性質を解明するための基礎的研究を行った．また，パラメタ化文字列照合問題と回文照合問題との関連性を明らかにし，回文照合問題を高速に解くアルゴリズムを開発した．

研究成果の概要（英文）：Parameterized string matching algorithms can be used to solve several important problems on natural language text searching and biological sequence pattern discovery. We proposed efficient algorithms to solve the parameterized string matching problem, and we revealed new properties of data structures for parameterized string matching. Also, we showed a close relationship between parameterized string matching and palindrome matching, and proposed efficient algorithms for palindrome matching.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,900,000	570,000	2,470,000
２０１０年度	900,000	270,000	1,170,000
年度			
年度			
年度			
総　計	2,800,000	840,000	3,640,000

研究分野：情報科学
科研費の分科・細目：情報学・情報学基礎
キーワード：アルゴリズム理論，文字列処理，データ構造，パタン発見

１．研究開始当初の背景

記憶媒体の低価格化やネットワークの高速化を背景として，計算機可読なデータが爆発的な速度で増加している．この膨大なデータからユーザが真に必要なデータを抽出するための技術の確立が急務となっている．本課題では，計算機可読なデータの多くが文字列データと見なせることに着目し，高速な文字列照合技法に関する研究を行う．

文字列ｓ中の文字を置き換えることで，文字列ｔと合致するとき，文字列ｓとｔはパラメタ化合致するという．パラメタ化文字列照合問題とは，パタン文字列ｐとテキスト文字列ｔが与えられたとき，ｐがｔ中でパラメタ化合致する位置の集合を求める問題である．パ

ラメタ化文字列照合は，ソフトウェアメンテナンスや盗作検出，RNA 配列の２次構造照合など，計算機科学やバイオ情報学の重要課題の基盤となるものである．しかしながら，素朴な方法では，パラメタ化文字列照合問題を解くのに $O(n^2\pi!)$時間を要してしまう．ここで，πは入力文字列中に現れる異なる文字の種類数である．
1996 年，Baker によってパラメタ化文字列照合問題を $O(n\log\pi)$時間・$O(n)$領域で解く画期的な手法が提案された．Baker のこの研究成果を契機に，パラメタ化文字列照合に関する研究が国内外で盛んに行われるようになっている．Baker のアルゴリズムは，パラメタ化接尾辞木というデータ構造を利用する．しかしながら，パラメタ化接尾辞木の領域計算量 $O(n)$に隠された定数項が非常に大きいことが知られている．したがって，大規模なテキストデータに適用することが困難である．

２．研究の目的

本研究では，計算機可読なデータの多くが文字列データと見なせることに着目し，文字列データに対するパタン照合技術の開発を行う．特に，パラメタ化文字列照合問題に注目し，これを高速かつ省領域で解くアルゴリズムの実現を目指す．具体的には，パラメタ化接尾辞木よりも省領域でパラメタ化文字列照合を解くことができるデータ構造についての研究開発を行う．

３．研究の方法

九州大学システム情報科学研究院および東北大学情報科学研究科のグループと連携して研究開発を行った．

４．研究成果

(1) パラメタ化接尾辞配列とパラメタ化 LCP 配列は，研究代表者らが世界に先駆けて 2008 年に提案した新しいデータ構造である．入力テキスト中の異なる文字の種類が高々２のとき，パラメタ化接尾辞配列とパラメタ化LCP配列を入力テキスト長の線形時間で構築するアルゴリズムが知られている．本研究課題では，テキストに現れる異なる文字の種類が２よりも大きいときに，パラメタ化接尾辞配列とパラメタ化LCP配列を高速に構築するアルゴリズムを開発した．これらの配列を用いることにより，パラメタ化文字列照合問題を $O(m+\log n+|Occ|))$時間で解くことができる．ここで，m はパタン文字列 p の長さ，n はテキスト文字列 t の長さ，Occ は p が t 中でパラメタ化合致する位置の集合である．提案手法が既存手法よりも高速にパラメタ化接尾辞配列とパラメタ化LCP配列を構築することを計算機実験によって確かめた．

(2) パラメタ化ボーダ配列とは，Morris と Pratt によって提案されたボーダ配列をパラメタ化文字列照合用に改良したデータ構造である．パラメタ化ボーダ配列を用いることによって，パラメタ化文字列照合問題を $O(n+|Occ|)$時間で解くことができる．本研究では，与えられた整数列をパラメタ化ボーダ配列として持つ文字列を出力する問題（逆問題）に世界に先駆けて取り組み，以下の研究成果を達成した．

(ア) アルファベットサイズが高々２のときに逆問題を線形時間で解くアルゴリズムを開発した．

(イ) アルファベットサイズが２より大きいとき，逆問題を $O(n^{1.5})$時間で解くアルゴリズムを開発した．

いずれの成果も，パラメタ化ボーダ配列に内在する様々な性質を解き明かし，それらを有機的に組み合わせることによって，アルゴリズムの高速化を達成したものである．提案アルゴリズムは，この逆問題を多項式時間で解く世界初の手法である．

(3) 回文照合問題とは，パタン文字列と回文構造が一致するテキスト文字列中の位置をすべて求める問題である．ここで，回文構造とは，各位置での極大な回文の長さのことをいう．
パラメタ化文字列照合と回文照合の関係性は，非自明である．しかしながら，研究代表者らは，パラメタ化照合問題と文字列中の回文構造が密接な関係を持

つことを解明した．具体的には，入力文字列中に現れる異なる文字の種類数が３以下のとき，文字列 p と s がパラメタ化合致することと，文字列 p と s の回文構造が一致することが同値であることを示した．このことによって，アルファベットサイズが３のときに，回文構造照合問題は線形時間で解けることを世界で初めて明らかにした．

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５件）

(1) Wataru Matsubara, Shunsuke Inenaga, and Ayumi Shinohara
An Efficient Algorithm to Test Square-Freeness of Strings Compressed by Balanced Straight Line Programs
Chicago Journal of Theoretical Computer Science, Special Issue: CATS 2009, Article 4, 2010.

(2) Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda
Counting and Verifying Maximal Palindromes
In Proc. the 17th Symposium on String Processing and Information Retrieval (SPIRE 2010), Lecture Notes in Computer Science 6393 (LNCS 6393), pp. 135-146, Springer-Verlag, 2010.

(3) Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda
Verifying a Parameterized Border Array in $O(n^{1.5})$ Time
In Proc. the 21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010), Lecture Notes in Computer Science (LNCS 6129), pp. 238-250, Springer-Verlag, 2010.

(4) Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda
Counting Parameterized Border Arrays for a Binary Alphabet
In Proc. 3rd International Conference on Language and Automata Theory and Applications (LATA 2009), Lecture Notes in Computer Science (LNCS 5457), pp. 422-433, Springer-Verlag, 2009.

(5) Tomohiro I, Satoshi Deguchi, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda
Lightweight Parameterized Suffix Array Construction
In Proc. 20th International Workshop on Combinatorial Algorithms (IWOCA 2009), Lecture Notes in Computer Science (LNCS 5874), pp. 312-323, Springer-Verlag, 2009.

〔学会発表〕（計４件）

(1) Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda
Counting Parameterized Border Arrays for a Binary Alphabet
コンピュテーション研究会
2010 年 9 月 29 日，長岡技術科学大学マルチメディアセンター

(2) 井 智弘, 稲永 俊介, 坂内 英夫, 竹田 正幸
Verifying a Parameterized Border Array in $O(n^{1.5})$ Time
第 9 回情報科学技術フォーラム（FIT 2010）
2010 年 9 月 7 日，九州大学伊都キャンパス

(3) 井 智弘, 稲永 俊介, 竹田 正幸
回文照合問題
冬のLAシンポジウム
2011 年 2 月 2 日，京都大学数理解析研究所

(4) 井 智弘, 出口 悟史, 坂内 英夫, 稲永 俊介, 竹田正幸
Lightweight Construction of Parameterized Suffix Arrays
夏のLAシンポジウム 2009
2009 年 7 月 23 日，宮城県東松島市

６．研究組織

(1)研究代表者

稲永　俊介（INENAGA SHUNSUKE）
九州大学・大学院システム情報科学
研究院・特任准教授
研究者番号：60448404