様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月２４日現在

研究種目：若手研究（B）
研究期間：2008～2009
課題番号：20700141
研究課題名（和文） 文字列パターン発見および文字列データ分類における
モデル選択アルゴリズムの研究

研究課題名（英文） Algorithms for Model Selection in String Pattern Discovery and String Data Classification

研究代表者
坂内 英夫（BANNAI HIDEO）
九州大学・大学院システム情報科学研究院・准教授
研究者番号：20323644

研究成果の概要（和文）：

本課題では与えられた大量の文字列データから，その特徴を捉えた意味のある文字列上のパターンを発見する問題，またはパターンに基づいてデータを分類する問題に対して，様々なパターンクラスの中から適切なパターンクラスを効率よく選択するための手法に関して研究を行った. 主な成果としては VLDC パターンクラス及びパラメータ化部分文字列パターンクラスに対して効率の良いアルゴリズムとデータ構造を提案し，その有用性を計算機実験によって示した.

研究成果の概要（英文）：

This study is concerned with developing efficient methods for selecting a good pattern class for pattern discovery from string data, as well as classification of string data based on patterns. We developed efficient algorithms and data structures for string pattern discovery and string data classification for several different classes of patterns, namely the VLDC pattern class, and the parameterized substring pattern class.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	**1,600,000**	**480,000**	**2,080,000**
２００９年度	**1,400,000**	**420,000**	**1,820,000**
年度			
年度			
年度			
総　計	**3,000,000**	**900,000**	**3,900,000**

研究分野：情報科学
科研費の分科・細目：情報学・知能情報学

キーワード：文字列アルゴリズム，文字列パターン発見

１．研究開始当初の背景

近年，インターネットの普及や様々な生物種のゲノム配列が決定されたこと等をはじめ，多岐にわたる分野で膨大な量の文字列データが生み出されており，利用可能となっている．与えられた大量の文字列データからデータの特徴を捉えた意味のあるパターンを効率的に発見する手法は，非常に応用性が高い重要な技術である．
文字列パターン発見においてそのパターンの探索範囲をパターンクラスと言い，最も単純な部分文字列パターンなど表現力が限られたものから，文字の不一致を許容した近似文字列パターン，可変長のワイルドカードを含む VLDC (Variable Length Don't Care) パターン，正規表現パターンなど，表現力が高いものまで様々なものを考えることができる．これらの中から問題に適した，適切なパターンクラスを用い，データの特徴を捉えたパターンを発見する問題，もしくはパターンを用いてデータを高精度に分類する問題を解くことは重要である．

２．研究の目的

本研究の目的は，様々あるパターンクラスから，与えられた文字列データの特徴を最も良く捉えられるパターンクラスもしくは分類精度が高くなるパターンクラスを効率よく探すための手法を開発することである．

３．研究の方法

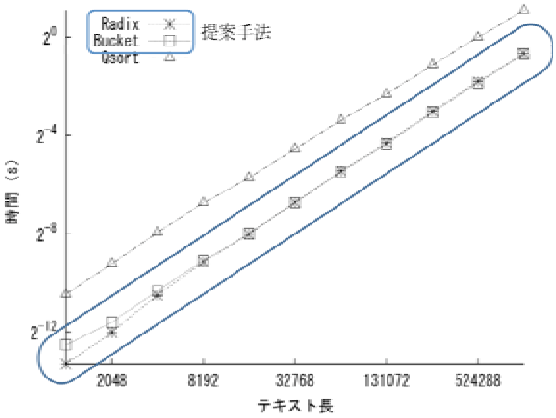
本研究では，新しいパターンクラスの提案，様々なパターンクラスにおいて効率良くパターン発見・データ分類扱うためのデータ構造の開発を行った．文字列データ分類には特にサポートベクトルマシンと文字列カーネルに着目し，パターンクラスによる分類はパターンクラスに基づく文字列カーネルを用いる．パターンクラスとしては，特に VLDC パターンクラス，パラメータ化部分文字列パタークンラスを扱う．

４．研究成果

本研究では主な成果は以下の通りである

(1) VLDC パターンクラスに基づく VLDC 文字列カーネルの提案
文字列パターンに基づく文字列カーネルとして，従来の部分列カーネルおよび部分文字列カーネルの特徴空間を含む，VLDC (Variable Length Don't Care) パターンに基づく VLDC カーネルを設計し，WDAWG (Wildcard DAWG) と呼ばれるデータ構造を用いてカーネル値を計算するアルゴリズムを開発した．計算機実験において，データによっては部分列カーネルよりも精度が良い場合があることを示した．

(2) パラメータ化文字列照合に基づく新たなパターンクラス:パラメータ化部分文字列パターンの提案，及びパターン発見･データ分類を効率よく行うためのアルゴリズム・データ構造の考案
パラメータ化文字列照合とは，アルファベットの各文字の入れ替えという曖昧性を許して照合するパターン照合であり，ソフトウェアにおける重複ソースコート検索，文書における剽窃発見，RNA の構造照合なと,様々な応用がある．本研究ではパラメータ化照合に基づくパターンクラスとしてパラメータ化部分文字列パターンクラスを新たに提案する．パラメータ化文字パターンの照合を効率良く行うため，従来はパラメータ化接尾辞木が用いられていたが，本研究ではより省メモリであるパラメータ化接尾辞配列 PSA を提案した．また，通常の接尾辞配列 SA では最長共通接頭辞配列 LCP を用いることで，接尾辞木上の操作が接尾辞配列で行えることが知られているが，パラメータ化接尾辞配列についても同様にパラメータ化最長共通接頭辞配列 PLCP を考えることでパラメータ化接尾辞木上の操作がパラメータ化接尾辞配列で行えることを示した．
これら二つの配列について，効率的な構築アルゴリズムを考案した．具体的には，バイナリアルファベット，つまり２種類の文字のみ



からなる文字列については，文字列の長さに対して最悪の場合でも線形時間で PSA および PLCP 配列を構築するアルゴリズムを示した．一方で，一般のアルファベットに対しては最悪計算量の理論値は文字列の長さ n に対して $O(n^2)$時間となってしまうものの，様々なデータに対して素朴なアルゴリズムよりも大幅に高速に PSA （上図）および PLCP を構築できるアルゴリズムを開発した．また，これらのデータ構造を合わせて用いることで，最適パラメータ化部分文字列パターン発見のアルゴリズム，及びパラメータ化部分文字列カーネルを高速に計算するアルゴリズムを実現した．

（３）繰り返し構造発見問題に関するビット並列アルゴリズムの考案

繰り返し構造は文字列上の基本的かつ重要な特徴である．本研究では連と呼ばれる繰り返し構造について，特に短い文字列に対して極めて高速に動作するビット並列アルゴリズムを開発した．従来の線形時間アルゴリズムは LZ 分解など複雑な操作が必要であったのに対し，提案手法は非常に簡潔であり，大量の文字列に対して処理を並列化することが容易である．実際にアルゴリズムを用いて長さ 48 以下のすべてのバイナリ文字列に対する繰り返し構造を列挙することができ，繰り返し構造の多い文字列の性質について幾つかの知見を得た

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 7 件）

1. Kazunori Hirashima, Hideo Bannai, Wataru Matsubara, Akira Ishino and Ayumi Shinohara,
“Bit-parallel algorithms for computing all the runs in a string”,
In Proceedings of The Prague Stringology Conference 2009 (PSC 2009), 203-213, 2009,
査読有.
2. Tomohiro I, Satoshi Deguchi, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda,
“Lightweight Parameterized Suffix Array Construction”,
In Proceedings of the 20th International Workshop on Combinatorial Algorithms (IWOCA 2009), LNCS 5874: 312-323, 2009.
査読有.
3. Wataru Matsubara, Kazuhiko Kusano, Hideo Bannai and Ayumi Shinohara,
“A Series of Run-rich Strings”,
In Proceedings of the 3rd International Conference on Language and Automata Theory and Applications (LATA 2009), LNCS 5457:578-587, 2009,
査読有.
4. Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda,
“Counting Parameterized Border Arrays for a Binary Alphabet”,
In Proceedings of the 3rd International Conference on Language and Automata Theory and Applications (LATA 2009), LNCS 5457:422-433, 2009,
査読有.
5. Kazuyuki Narisawa, Hideo Bannai, Kohei Hatano, Shunsuke Inenaga, Masayuki Takeda,
“String Kernels Based on Variable-Length-Don't-Care Patterns”,
In Proceedings of the 11th International Conference on Discovery Science (DS2008), LNAI 5255:308-318,2008,
査読有.
6. Satoshi Deguchi, Fumihito Higashijima, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda,
“Parameterized Suffix Arrays for Binary Strings”,
In Proceedings of The Prague Stringology Conference 2008 (PSC2008), 84-94, 2008,
査読有.
7. Wataru Matsubara, Kazuhiko Kusano, Akira Ishino, Hideo Bannai and Ayumi Shinohara,
“New Lower Bounds for the Maximum Number of Runs in a String”,
In Proceedings of The Prague Stringology Conference 2008 (PSC2008), 140-145, 2008,
査読有.

〔学会発表〕（計 8 件）

1. Kazunori Hirashima, Hideo Bannai, Wataru Matsubara, Akira Ishino and Ayumi Shinohara, “Bit-parallel algorithms for computing all the runs in a string”, Prague Stringology Conference 2009, 2009 年 9 月 2 日, プラハ, チェコ共和国.
2. 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.

3. Tomohiro I, Satoshi Deguchi, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, “Lightweight Parameterized Suffix Array Construction”, 20th International Workshop on Combinatorial Algorithms, 2009年6月29日, Hradec nad Moravicí, チェコ共和国.
4. Wataru Matsubara, Kazuhiko Kusano, Hideo Bannai and Ayumi Shinohara,“A Series of Run-rich Strings”, 3rd International Conference on Language and Automata Theory and Applications, 2009 年 4 月 7 日，タラゴナ，スペイン.
5. Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, “Counting Parameterized Border Arrays for a Binary Alphabet”, 3rd International Conference on Language and Automata Theory and Applications, 2009 年 4 月 2 日, タラゴナ, スペイン.
6. Kazuyuki Narisawa, Hideo Bannai, Kohei Hatano, Shunsuke Inenaga, Masayuki Takeda, “String Kernels Based on Variable-Length-Don't-Care Patterns”, International Conference on Discovery Science (DS2008), 2008年10月14日，ブダペスト,ハンガリー.
7. Wataru Matsubara, Kazuhiko Kusano, Akira Ishino, Hideo Bannai and Ayumi Shinohara, “New Lower Bounds for the Maximum Number of Runs in a String”, Prague Stringology Conference 2008, 2008 年 9 月 3 日，プラハ，チェコ共和国.
8. Satoshi Deguchi, Fumihito Higashijima, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, “Parameterized Suffix Arrays for Binary Strings”, Prague Stringology Conference 2008, 2008 年 9 月 1 日，プラハ，チェコ共和国.

６．研究組織

(1)研究代表者

坂内英夫（BANNAI HIDEO）

九州大学・大学院システム情報科学研究院・准教授

研究者番号：20323644

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：