様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ９ 日現在

機関番号：１１３０１
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０２０８

研究課題名（和文）様々なパターン照合問題に対する時間・空間計算量的に最適なアルゴリズムの開発

研究課題名（英文）Development of optimal time-space algorithms on pattern matching problems

研究代表者
ＨＥＮＤＲＩＡＮ　ＤＩＰＴＡＲＡＭＡ（HENDRIAN, DIPTARAMA）

東北大学・情報科学研究科・助教

研究者番号：７０８２３１３６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：パターン照合問題はテキスト文字列にパターン文字列の出現位置を求める問題である．パターン照合アルゴリズムは文字列検索やデータマイニングに応用され，データ処理において基本的な技術である．本研究は厳密パターン照合問題およびその拡張であるパラメータ照合問題や順序保存パターン照合問題に対して様々な線形時間アルゴリズムおよび索引構造を提案した．さらに，パラメタ化パターン照合に対して劣線形領域アルゴリズムを開発し，正当性および計算量の証明することができた．本アルゴルズムは Galil& Seiferas による厳密パターン照合アルゴリズムを拡張したアルゴリズムである．

研究成果の学術的意義や社会的意義
本研究はパラメタ化パターン照合問題に対して，初めて劣線形領域アルゴリズムを提案した．劣線形領域アルゴリズムは情報科学理論において学術的に非常に重要である．また，パターン照合アルゴリズムは文字列検索を使ったアプリケーションに応用されるとともに，バイオインフォマティクスや人工知能等，様々な分野の研究に用いられ，非常に重要な技術である．そのため，本研究の成果は社会のおける情報科学技術の発展およびこれからの研究の発展に貢献できたと考えられる，

研究成果の概要（英文）：The pattern matching problem is a task to find occurrences of pattern strings in text strings. Pattern matching algorithms are fundamental in computer science since pattern matching algorithms have broad applications such as data searching and data mining. In this study, we consider the exact pattern matching problem and its extension, such as parameterized pattern matching and order-preserving pattern matching problems. We proposed various linear time pattern matching algorithms and indexing structures for the exact pattern matching problem and its extension. Moreover, we designed a sub-linear extra-space algorithm for the parameterized pattern matching problem. Our algorithm is based on the Galil&Sefeiras sub-linear extra-space algorithm for exact pattern matching.

研究分野：文字列学，パターン検索

キーワード：パターン照合アルゴリズム　索引構造　パラメタ化パターン照合　順序保存パターン照合　劣線形領域アルゴリズム　アルゴリズム理論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

パターン照合問題はデータ解析において重要な基本問題である．パターン照合（厳密パターン照合と呼ぶ)はテキスト中にパターンと厳密に一致する部分文字列を求める問題である.また，応用例を考えると厳密パターン照合だけではなく，近似的なパターン照合やデータの特徴に特化したパターン照合も必要であり，データの特徴に合わせたパターン照合（拡張したパターン照合と呼ぶ，図１参照）アルゴリズムが提案されてきた．データ解析を効率的に行うために高速なアルゴリズムだけではなく，小型ロボットや宇宙探査機といった容量が非常に限られたシステムにおいて,省メモリなアルゴリズムも非常に重要である.厳密パターン照合問題に対して，様々なアルゴリズムが提案されてきた．その中には，定数領域かつ線形時間という，計算量の観点から最適な領域と時間でパターン照合問題を解くものも既に存在する．しかし，拡張パターン照合問題に対して文字列の長さより少ない作業領域で動作する劣線形領域アルゴリズムはまだ提案されていない.

２．研究の目的

一般的には拡張したパターン照合問題は厳密パターン照合問題より困難であることが知られている．なぜならば，拡張した文字列の構造によって，通常の文字列には成り立つ法則が拡張した文字列に成り立たないことがあるからである．たとえば，一般的な文字列において，任意の周期についてその周期の倍数も周期であることがよく知られており，この性質を用いて最適なパターン照合アルゴリズムが開発されてきた．しかし，この性質はパラメタ化文字列において必ずしも成立するわけではない．そのため，厳密パターン照合アルゴリズムを拡張パターン照合アルゴリズムに適用できるかどうかは自明ではない．そこで，本研究は，拡張したパターン照合問題において，各問題で扱われる文字列の構造を解明しながら，それを活用することで，線形時間かつ劣線形領域アルゴリズムを開発する．

３．研究の方法

パターン照合を高速に行うためにはパターンおよびテキストの性質が非常に重要である．特にパターン照合やデータ圧縮によく使われる性質である「繰り返し構造」に着目する．省メモリなアルゴリズムに有用な着想を得るために，各文字列の繰り返し構造を中心に拡張した文字列の性質を理論的に解析するとともに，様々なデータを用いて計算機実験を行う．次に，解明した拡張パターン照合問題で扱われる文字列の性質を活用して，拡張パターン照合の線形時間・線形領域アルゴリズムや索引構造を開発する．さらに，拡張したパターン照合問題に対して文字列の長さより少ない領域で動作する線形時間・劣線形領域アルゴリズムを開発する．

４．研究成果

(1) 厳密パターン照合問題に対する線形時間アルゴリズム

本研究において厳密パターン照合問題に対する線形時間アルゴリズムに関する成果を 2 つあげられた．１つ目は既存のアルゴリズムである Franek-James-Smith アルゴリズムを改良したアルゴリズムである．提案アルゴリズムは理論的に線形時間で動作し，実験的に既存の Franek-James-Smith アルゴリズムより高速に動作することを示した．2 つ目のアルゴリズムはパターンにおける q-グラムの出現位置の距離を用いた線形時間かつ実験的に高速なアルゴリズムである．一部のデータセットで提案アルゴリズムは既存アルゴリズムより速いことを実験で示した．

(2) 厳密パターン照合問題に対する索引構造のオンライン構築アルゴリズム

パターン照合問題に対する索引構造である線形領域接尾辞トライに対するオンライン構築アルゴリズムを提案した．本論文ではテキスト文字列を左から入力される場合のアルゴリズムおよび左から入力されたアルゴリズム，２種類のアルゴリズムを提案した.また，索引構造として使われる Burrows-Wheeler 変換の一種である Bijective Burrows-Wheeler 変換を定数作業領域で行うアルゴリズムを提案した．本アルゴリズムは入力の文字列の領域をそのまま使って出力の文字列に書き換えることで，省メモリで動作することができる．

(3) 拡張パターン照合問題に対する線形時間アルゴリズムおよび索引構造の構築

まず，マルチトラック文字列に対して，線形時間アルゴリズムや実験的に高速なアルゴリズム等，様々な順列パターン照合アルゴリズムを提案した．次にパラメタ化パターン照合問題に対してパラメタ化線形領域接尾辞トライ，パラメタ化有向無閉路文字列グラフといった索引構造とその構築アルゴリズムを提案した．また，索引構造であるパラメタ化 Burrows-Wheeler 変換（PBWT 変換）のオンラインアルゴリズムを提案した．

(4) 拡張パターン照合問題に対する並列アルゴリズム

パラメタ化一致や順序同形などを含んだ部分文字列一貫同値関係クラス上のパターン照合問題に対して，並列アルゴリズムを提案した．提案アルゴリズムを用いて，部分文字列一貫同値関係クラスに含まれた任意同値関係上のパターン照合問題を解くことができる．

(5) パラメタ化パターン照合問題に対する劣線形領域アルゴリズム

パラメタ化パターン照合に対して劣線形領域アルゴリズムを開発し，正当性および計算量の証明することができた．本アルゴルズムは Galil&Seiferas による劣線形領域の厳密パターン照合アルゴリズムを拡張したアルゴリズムである．

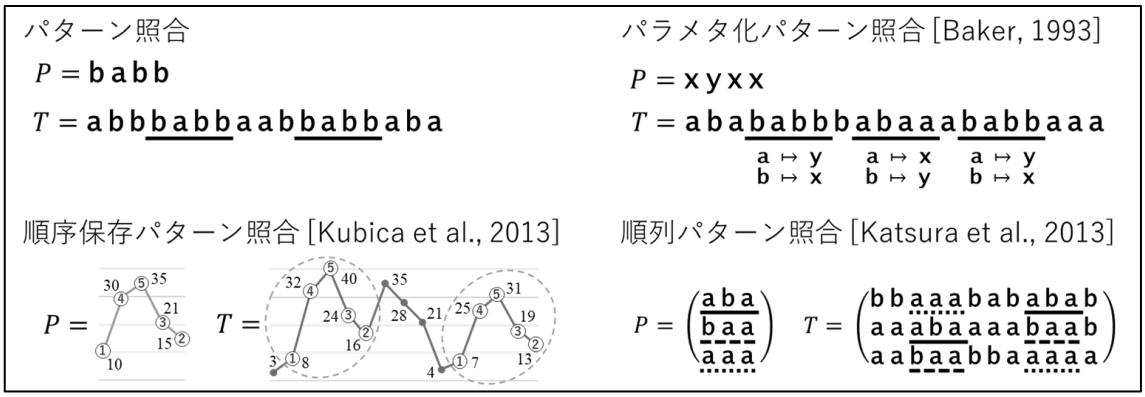

図 1 パターン照合とその拡張の例

５．主な発表論文等

〔雑誌論文〕　計15件（うち査読付論文　15件／うち国際共著　0件／うちオープンアクセス　9件）

１．著者名 Koshiro Kumagai, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	４．巻 13973
２．論文標題 Inferring Strings from Position Heaps in Linear Time	５．発行年 2023年
３．雑誌名 Proceedings of the 17th International Conference and Workshops on Algorithms and Computation	６．最初と最後の頁 115-126
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-27051-2_11	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nakashima Katsuhito, Fujisato Noriki, Hendrian Diptarama, Nakashima Yuto, Yoshinaka Ryo, Inenaga Shunsuke, Bannai Hideo, Shinohara Ayumi, Takeda Masayuki	４．巻 933
２．論文標題 Parameterized DAWGs: Efficient constructions and bidirectional pattern searches	５．発行年 2022年
３．雑誌名 Theoretical Computer Science	６．最初と最後の頁 21-42
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.tcs.2022.09.008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Daiki Hashimoto, Diptarama Hendrian, Dominik Koppl, Ryo Yoshinaka, Ayumi Shinohara	４．巻 13617
２．論文標題 Computing the Parameterized Burrows Wheeler Transform Online	５．発行年 2022年
３．雑誌名 Proceedings of the 29th International Symposium on String Processing and Information Retrieval	６．最初と最後の頁 70-85
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-20643-6_6	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Davaajav Jargalsaikhan, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	４．巻 223
２．論文標題 Parallel Algorithm for Pattern Matching Problems Under Substring Consistent Equivalence Relations	５．発行年 2022年
３．雑誌名 Proceedings of the 33rd Annual Symposium on Combinatorial Pattern Matching	６．最初と最後の頁 28:1-21
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.CPM.2022.28	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Natsumi Kikuchi, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	4．巻 12303
2．論文標題 Computing Covers Under Substring Consistent Equivalence Relations	5．発行年 2020年
3．雑誌名 Proceedings of the 27th International Symposium on String Processing and Information Retrieval, Lecture Notes in Computer Science	6．最初と最後の頁 131-146
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-59212-7_10	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Dominik Koppl, Daiki Hashimoto, Diptarama Hendrian, Ayumi Shinohara	4．巻 161
2．論文標題 In-Place Bijective Burrows-Wheeler Transforms	5．発行年 2020年
3．雑誌名 Proceedings of the 31st Annual Symposium on Combinatorial Pattern Matching, Leibniz International Proceedings in Informatics	6．最初と最後の頁 21:1-15
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.CPM.2020.21	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Satoshi Kobayashi, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	4．巻 160
2．論文標題 Fast and linear-time string matching algorithms based on the distances of q-gram occurrences	5．発行年 2020年
3．雑誌名 Proceedings of 18th Symposium on Experimental Algorithms, Leibniz International Proceedings in Informatics	6．最初と最後の頁 13:1-13
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.SEA.2020.13	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Hendrian Diptarama	4．巻 12049
2．論文標題 Generalized Dictionary Matching Under Substring Consistent Equivalence Relations	5．発行年 2020年
3．雑誌名 Proceedings of the 14th International Conference and Workshop on Algorithms and Computation, Lecture Notes in Computer Science	6．最初と最後の頁 120-132
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-39881-1_11	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Ryu Wakimoto, Satoshi Kobayashi, Yuki Igarashi, Davaajav Jargalsaikhan, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	4．巻 2568
2．論文標題 AOBA: An Online Benchmark tool for Algorithms in stringology	5．発行年 2020年
3．雑誌名 Proceedings of the SOFSEM 2020 Student Research Forum, CEUR Workshop Proceedings	6．最初と最後の頁 1-12
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Katsuhito Nakashima, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	4．巻 2568
2．論文標題 An Extension of Linear-size Suffix Tries for Parameterized Strings	5．発行年 2020年
3．雑誌名 Proceedings of the SOFSEM 2020 Student Research Forum, CEUR Workshop Proceedings	6．最初と最後の頁 97-108
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Satoshi Kobayashi, Diptarama Hendrian, Ryo Yoshinaka, Ayumi Shinohara	4．巻 -
2．論文標題 An improvement of the Franek-Jennings-Smyth pattern matching algorithm	5．発行年 2019年
3．雑誌名 Proceedings of the Prague Stringology Conference 2019	6．最初と最後の頁 56-68
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Diptarama Hendrian, Takuya Takagi, Shunsuke Inenaga	4．巻 128
2．論文標題 Online Algorithms for Constructing Linear-size Suffix Trie	5．発行年 2019年
3．雑誌名 Proceedings of the 30th Annual Symposium on Combinatorial Pattern Matching, Leibniz International Proceedings in Informatics	6．最初と最後の頁 30:1-19
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.CPM.2019.30	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Diptarama Hendrian, Yohei Ueki, Kazuyuki Narisawa, Ryo Yoshinaka, Ayumi Shinohara	4．巻 12
2．論文標題 Permuted Pattern Matching Algorithms on Multi-Track Strings	5．発行年 2019年
3．雑誌名 Algorithms	6．最初と最後の頁 73:1-20
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a12040073	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 Dominik Koppl, Daiki Hashimoto, Diptarama Hendrian, Ayumi Shinohara
2．発表標題 In-Place Bijective Burrows Wheeler Transformations
3．学会等名 Workshop Data Structures in Bioinformatics（国際学会）
4．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関