氏名（本籍地）	DIPTARAMA（ディプタラマ） HENDRIAN（ヘンリアン）
学位の種類	博　士（情報科学）
学位記番号	情　博　第　656　号
学位授与年月日	平成３０年　３月２７日
学位授与の要件	学位規則第４条第１項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Fast Algorithms on Pattern Matching Problems （パターン照合問題に対する高速なアルゴリズム）
論文審査委員	（主査）東北大学教　授　篠原　歩 東北大学教　授　徳山　豪　　東北大学教　授　周　暁 東北大学准教授　吉仲　亮

論文内容の要旨

第１章 序論

パターン照合問題はテキスト文字列中にパターン文字列の出現位置を求めるものであり，文字列処理における最も基本的な要素技術の一つである．その発展形として，実用性や応用例，データの形式等に即して様々な問題設定がなされてきた．例えば複数のパターンをテキストの中から同時に探し出す辞書照合や，プログラムソースコードの盗用検出を動機として変数名の付け替えを許容したパラメータ化パターン照合については，長い研究の歴史がある．また近年では，数値列データの相対的な大小関係のみに着目した順序保存パターン照合や，複数の文字列からなるマルチトラック文字列に対する順列パターン照合など，新たな照合問題に対する関心も高まっている．

本論文では上述の様々な問題設定において，パターン照合を高速に行うための新たなデータ構造やアルゴリズムを提案する．計算量の解析と実行時間の測定によって，理論と応用の両面からアルゴリズムの性能を評価する．表1は提案アルゴリズムの一覧を示す．

表 1 各問題に対する提案アルゴリズムの一覧

問題	アルゴリズム
半動的な辞書照合	有向無閉路文字列グラフを用いたアルゴリズム
	Aho-Corasick オートマトンを用いたアルゴリズム
動的な辞書照合	有向無閉路文字列グラフを用いたアルゴリズム
	Aho-Corasick オートマトンを用いたアルゴリズム
パラメータ化パターン照合	パラメータ化ポジションヒープ
順序保存パターン照合	Duel-and-sweep アルゴリズム
順列パターン照合	マルチトラック Knuth-Morris-Pratt アルゴリズム
	マルチトラック Aho-Corasick オートマトン
	マルチトラック順列パターン照合オートマトン
	マルチトラック Boyer-Moore アルゴリズム
	マルチトラック Horspool アルゴリズム
	フィルタを用いたアルゴリズム
	マルチトラック duel-and-sweep アルゴリズム

第２章 準備

第２章では，文字列に関する用語や表記の定義を説明する．記号の集合をアルファベット Σ と

呼ぶ. 文字列 X について, X の長さを $|X|$ で表す. 任意の位置 $i\,(1 \le i \le |X|)$ に対して $X[i]$ を i 番目の文字とする，また，任意の位置 $i,j\,(1 \le i \le j \le |X|)$ に対して，$X[i{:}j]$を位置 i から位置 j までの X の部分文字列とする.

また，本研究と関連する既存アルゴリズムである Knuth-Morris-Pratt アルゴリズム，Boyer-Moore アルゴリズム，Horspool アルゴリズム，および duel-and-sweep アルゴリズムについて説明する.

第 3 章 データ構造

第3章では, 本論文で提案されるアルゴリズムやデータ構造と関連したデータ構造について説明する. 具体的には Aho-Corasick オートマトン，有向無閉路文字列グラフ, およびポジションヒープに関する説明を行う.

第 4 章 動的な辞書照合アルゴリズム

辞書問題は辞書と呼ばれるパターンの集合とテキストを与えられ, テキスト中にすべてのパターンの出現位置を求める問題である. 動的な辞書照合問題は辞書照合の拡張であり, 辞書に対して新しいパターンの挿入および辞書中のパターンの削除を許した辞書照合問題である. また, パターンの挿入のみを許した問題を半動的な辞書問題と呼ぶ.

半動的な辞書照合問題および動的な辞書照合問題, それぞれの問題に対するアルゴリズムを 2 つ提案する. 更新時間の計算量に重きを置いた 1 つ目のアルゴリズムは有向無閉路文字列グラフを用いたもので, 既存の同種のアルゴリズムと同じ照合時間を保ちながら, 更新時間はより高速である. 一方, 2 つ目のアルゴリズムは Aho-Corasick オートマトンを用いており, 更新時間はパターンによっては遅くなることがあるが, 照合時間は速く, 実用的なアルゴリズムである.

辞書に挿入されるまたは辞書から削除されるパターンを P , テキストを T , 辞書中のパターンの長さの総和を d , occ をテキスト中にパターンの出現数とする. 有向無閉路文字列グラフを用いたアルゴリズムは半動的な辞書照合問題において更新を $O(|P|\log|\Sigma|)$ 時間, 照合を $O(|T|\log|\Sigma| + occ)$ 時間で行うことができ, 動的な辞書照合問題において更新を $O(|P||\Sigma| + \log d/\log\log d)$ 時間, 照合を $O(|T|(\log d/\log\log d + \log|\Sigma|) + occ\log d/\log\log d)$ 時間で行うことができる. それに対して, Aho-Corasick オートマトンを用いたアルゴリズムは半動的な辞書照合問題において更新を $O(|P|\log|\Sigma| + u)$ 時間, 照合を $O(|T|\log|\Sigma| + occ)$ 時間で行うことができ, 動的な辞書照合問題において更新を $O(|P||\Sigma| + u)$ 時間, 照合を $O(|T|\log|\Sigma| + occ)$ 時間で行うことができる. ここで u は Aho-Corasick オートマトンにおいて更新される状態の数である.

第 5 章 パラメータ化パターン照合アルゴリズム

パラメータ化文字列とは定数記号と変数記号からなる文字列であり, 2つ同じ長さのパラメータ化文字列 X と Y に対して $f(X[i]) = Y[i]\,(1 \le i \le |X|)$ および定数記号に恒等写像となる単射 f が存在するとき, S は T にパラメータ化一致という. また, テキスト T 中にパターン P とパラメータ化一致となる部分文字列を求める問題をパラメータ化パターン照合問題と呼ぶ.

本章では文字列の索引構造の一つであるポジションヒープをパラメータ化パターン照合に適合するように拡張し, パラメータ化ポジションヒープを提案する. また, パラメータ化ポジションヒープの $O(|T|\log(|\Sigma| + |\Pi|))$ 時間構築アルゴリズムを提案する. ここで, Σ は定数アルファベット, Π は変数アルファベットである. さらに, 補助データ構造を用いて, $O(|P|\log(|\Sigma| + |\Pi|) + |P||\Pi| + occ)$ 時間照合アルゴリズムを提案する. ここで, occ はテキスト中にパターンの出現数である.

第 6 章 順序保存パターン照合アルゴリズム

同じ長さの整数文字列 X と Y に対して, 任意の $i,j\,(1 \le i,j \le |X|)$ で $X[i] \le X[j] \Leftrightarrow Y[i] \le Y[j]$ が成り立つとき，X と Y は順序同型であるという. 順序保存パターン照合問題とは与えられたテキスト T の中にパターン P と順序同型となる部分文字列を求める問題である.

順序保存パターン照合問題に対して, duel-and-sweep アルゴリズムをもとにした新たなアルゴリズムを提案する. このアルゴリズムは$O(|P| + |T|)$ 時間で順序保存パターン照合を行うことができる. 本アルゴリ

ズムの理論計算量は既存の最良のものと同等であるが，実用上はより高速に動作することを計算機実験によって実証する．

第 7 章 順列パターン照合アルゴリズム

長さ n の文字列の k 項組をマルチトラック文字列 $\mathbb{X} = (X_1, X_2, \ldots, X_k)$ と呼び，省略してマルチトラックと呼ぶ．任意の $(1, \ldots, k)$ の順列 $r = (r_1, r_2, \ldots, r_k)$ に対して $\mathbb{X}\langle r\rangle = (X_{r_1}, X_{r_2}, \ldots, X_{r_k})$ を順列マルチトラックと呼ぶ．同じ長さのマルチトラック $\mathbb{X}$ と $\mathbb{Y}$ に対して $\mathbb{X} = \mathbb{Y}\langle r\rangle$ となる順列 r が存在するとき，$\mathbb{X}$ は $\mathbb{Y}$ と順列一致するという．順列パターン照合問題とはマルチトラックテキスト $\mathbb{T}$ とパターン $\mathbb{P}$ を与えられ，$\mathbb{P}$ と順列一致となる $\mathbb{T}$ の部分文字列を求める問題である．

マルチトラック文字列上の順列パターン照合問題に対する 7 種類のアルゴリズムを提案する．最初に Knuth-Morris-Pratt アルゴリズムをもとにしたアルゴリズムを提案し，さらにこれを発展させてマルチトラックオートマトンと順列パターン照合オートマトンという 2 つのデータ構造を提案する．マルチトラックオートマトンを用いたアルゴリズムは複数のパターンを同時に照合することができる．一方，順列パターン照合オートマトンは理論的にも実験的にも高速なアルゴリズムである．次に，パターンを右端から照合する Boyer-Moore アルゴリズムおよび Horspool アルゴリズムをもとにしたアルゴリズムを提案し，これらのアルゴリズムは理論的にはマルチトラック KMP アルゴリズムより遅いが，実験的には速いことを示す．次のアルゴリズムはフィルタを用いたものであり，与えられたテキストに対して高速な処理でパターンの出現位置の候補を出し，この候補に対してさらに検証を行う．フィルタを用いたアルゴリズムは候補を高速な処理で絞り込むものであり，候補が少ない文字列に対して非常に高速に動作するアルゴリズムである．最後は duel-and-sweep アルゴリズムをもとにしたアルゴリズムであり，このアルゴリズムは理論的に速いことを示す．表 2 は各提案アルゴリズムの計算量を示す．ここで d は辞書中のパターンの長さの総和である．

表 2 提案順列パターン照合アルゴリズムの計算量

アルゴリズム	前処理時間	照合時間
MT Knuth-Morris-Pratt アルゴリズム	$O(\|\mathbb{P}\|k)$	$O(\|\mathrm{T}\|k)$
MT Aho-Corasick オートマトン	$O(dk\log\|\Sigma\|)$	$O(\|\mathrm{T}\|k\log\|\Sigma\|)$
MT 順列パターン照合オートマトン	$O(\|\mathbb{P}\|k\log\|\Sigma\|)$	$O(\|\mathrm{T}\|k\log\|\Sigma\|)$
MT Boyer-Moore アルゴリズム	$O\big(\|\mathbb{P}\|(k\log\|\Sigma\| + \|\Sigma\|)\big)$	$O\big(\|\mathrm{T}\|k(\|\mathbb{P}\| + \log\|\Sigma\| + \|\Sigma\|)\big)$
MT Horspool アルゴリズム	$O\big(\|\mathbb{P}\|(k\log\|\Sigma\| + \|\Sigma\|)\big)$	$O\big(\|\mathrm{T}\|k(\|\mathbb{P}\| + \log\|\Sigma\| + \|\Sigma\|)\big)$
MT duel-and-sweep アルゴリズム	$O(\|\mathbb{P}\|k)$	$O(\|\mathrm{T}\|k)$

第 8 章 結論と今後の課題

第7章は結論であり，論文の成果のまとめおよび今後の研究課題について述べる．本論文では動的な辞書照合問題，パラメータ化パターン照合問題，順序保存パターン照合問題，および順列パターン照合問題に対するアルゴリズムとデータ構造を提案した．提案したアルゴリズムに対して計算量の解析で理論的に高速であること示し，計算機実験で実用的に高速であることを示した．また，今後の課題として，より高速なアルゴリズムの開発やパターン照合問題に対するアルゴリズムを発展問題への拡張があげられる．

論文審査結果の要旨

与えられたテキストの中から目的とするパターンの出現を探し出すパターン照合は，文字列処理における最も基本的な要素技術の一つである．その発展形として，実用性や応用例，データの形式等に即して様々な問題設定がなされてきた．例えば複数のパターンをテキストの中から同時に探し出す辞書照合や，プログラムソースコードの盗用検出を動機として変数名の付け替えを許容したパラメータ化パターン照合については，長い研究の歴史がある．また近年では，数値列データの相対的な大小関係のみに着目した順序保存パターン照合や，複数の文字列からなるマルチトラック文字列に対する順列パターン照合など，新たな照合問題に対する関心も高まっている．

著者は，上述の様々な問題設定において，パターン照合を高速に行うための新たなデータ構造やアルゴリズムの開発と，実行時間の解析に理論と応用の両面から取り組んできた．本論文はその成果をまとめたもので，全編８章からなる．

第 1 章は序論である．

第 2 章では，準備として論文中で用いる用語や表記を定義し，基礎的なパターン照合アルゴリズムを紹介している．

第 3 章では，本論文で使用する基本的なデータ構造を説明している．

第 4 章では，複数のパターンを一瞥で照合できる Aho-Corasick パターン照合オートマトンに対して，パターンの追加や削除を効率よく行う更新アルゴリズムを示している．

第 5 章では，文字列の索引構造の一つであるポジションヒープをパラメータ化パターン照合に適合するように拡張し，その構築アルゴリズムと照合アルゴリズムを示している．この手法により構築も照合もそれぞれテキストとパターンの線形時間で行えることが証明されている．

第 6 章では，順序保存パターン照合を行う新たなアルゴリズムを提案している．このアルゴリズムの理論計算量は既存の最良のものと同等であるが，実用上はより高速に動作することが計算機実験によって実証されている．

第 7 章では，マルチトラック文字列上の順列パターン照合問題に対する 7 種類のアルゴリズムを提案しており，そのうちの一つは複数のパターンを同時に照合するものである．それぞれについての計算量の解析がなされ，実用上の速度比較も行われている．またこの実験により，提案手法は既存のどの手法よりも高速であることが確かめられている．

第 8 章は結論であり，論文の成果をまとめると共に，今後の研究課題について述べている．

以上要するに本論文は，様々なパターン照合問題に対する高速なアルゴリズムとデータ構造について研究したものであり，文字列学や計算量理論を中心にシステム情報科学の発展に寄与するところが少なくない．

よって，本論文は博士（情報科学）の学位論文として合格と認める．