博 士（情報科学） 金 田 悠 作

学 位 論 文 題 名

Efficient Regular Expression Matching Algorithms and Their Implementation on Reconfigurable Hardware

（効率良い正規表現照合アルゴリズムとその再構成可能ハードウェア上の実装）

学位論文内容の要旨

パターン照合問題 (pattern matching problem) とは，あるパターン P とあるテキスト T が与えられたとき，P の T 中の出現位置をすべて出力する問題である．理論計算機科学分野では，これまでに文字列照合や，正規表現照合，木パターン照合等，さまざまなパターン照合問題が研究されている．近年のセンサー技術とネットワーク技術の性能向上に伴う大規模なデータストリームの増加により，これらのデータからの効率的な情報検索と情報発見が重要になってきた．本研究では，とくに，高速な大規模データストリームに対し，大量かつ複雑なパターンを照合する問題，すなわち，大規模パターン照合問題 (large-scale pattern matching problem) について考察する．大規模パターン照合は，大規模データストリーム処理の基盤技術であり，理論的観点からだけでなく，実用的観点からも重要である．

1992 年に Wu と Manber, および，Baeza-Yates と Gonnet が提案した Shift-And 手法は，入力として与えられたパターンを受理する非決定性有限オートマトン (nondeterministic finite automaton, NFA) の状態集合をビット列で表現し，論理演算と，シフト演算，表引き計算の計算機ワード内の並列性を利用することで，文字列に対するパターン照合問題を効率良く解く手法である．このような計算機ワード内の並列性を用いたアルゴリズムは，ビット並列アルゴリズム (bit-parallel algorithm) と呼ばれ，文字列処理分野を中心に広く研究されている．Shift-And 手法は，正規表現照合問題 (regular expression matching problem) に拡張できる．ここで，正規表現とは，文字列パターンの集合を表現し，文字と，連結，和，繰り返しから再帰的に定義される．一方で，正規表現の場合，NFA の空遷移閉包計算のために大規模な遷移表を用いるため，計算領域が大きい．

本研究では，2001 年に Navarro と Raffinot が提案した拡張 Shift-And 手法の考えに基づき，より複雑なパターンのクラスに対して，算術演算を用いる効率良いビット並列パターン照合アルゴリズムを設計する．また，ビット並列アルゴリズムは計算機ワードに対する論理演算と算術演算のみから構成されるため，条件分岐や主記憶領域の間接参照を多用するアルゴリズムと比較して，回路実装に適していると考えられる．そこで，本研究では，ビット並列アルゴリズムの再構成可能ハードウェア (reconfigurable hardware) 上の実装についても考察する．

初めに，第 2 章で基本的な概念と定義を準備した後で，第 3 章では，ネットワーク表現と正規表現のクラスに対するパターン照合問題ついて考察する．ここで，ネットワーク表現とは，文字列と，連結，和から定義される正規表現の部分クラスである．本章では，算術演算を用いたビット並列演算 scatter と gather を提案し，これらの演算を拡張 Shift-And 手法で用いられるビット並列演算 propagate と組み合わせることで，ネットワーク表現照合問題を $O(ndm/w)$ 時間と $O(dm/w)$ 領域で解くビット並列アルゴリズムを与える．ここで，m と n はパターン長とテキスト長であり，d はパターンに対する構文木の深さ，w は計算機ワード長である．また，ネットワーク表現に対するビット

並列アルゴリズムと monotone routing 技法を組み合わせることで, 一般の正規表現のクラスに対するパターン照合問題を $O(ndm\log(m)/w)$ 時間と $O(dm\log(m)/w)$ 領域で解くビット並列アルゴリズムを与える. ここで, m と n はパターン長とテキスト長であり, d はパターンに対する構文木の深さ, w は計算機ワード長である. また, ワード上のビット逆転を定数時間で実行可能なハードウェア上で, 上記の計算量が $O(ndm/w)$ 時間と $O(dm/w)$ 領域に改善できることを示す.

次に, 第 4 章では, 無順序木のクラスに対するパターン照合問題を考察する. 木パターン照合問題とは, あるパターン木 P とあるテキスト木 T が与えられたとき, P の T 中のすべての出現位置を出力する問題である. ここで, P が T 中のある節点 v で出現するとは, ある P の節点から T の節点への写像が存在し, P のすべての節点が T の節点に写像され, かつ P の根が v に写像されることをいう. 本章では, とくに, P の節点から T の節点への写像として多対一写像を許した木パターン照合問題である, 無順序木に対する擬似木パターン照合問題 (unordered pseudo-tree matching problem), および, 無順序木に対する木同相写像問題 (unordered tree homeomorphism problem) を考える. これらの問題は, XPath の検索問題と深く関連する. 一方で, その実用面での重要性にもかかわらず, 理論計算機科学分野における従来の理論的解析は, 一対一写像に関するものがほとんどであった. したがって, このような多対一写像を許した木パターン照合問題を考察することは, 実用的観点と理論的観点の両方から非常に重要である. 本章では, はじめに, 無順序木に対するパターン照合問題を $O(nm)$ 時間と $O(hm)$ 領域で解くアルゴリズムを与える. 次に, 算術演算を用いたビット並列演算 tree aggregation を与え, この演算を上述のアルゴリズム, および, Myers のモジュール分解技法と組み合わせることで, 無順序木に対するパターン照合問題を $O(nm\log(w)/w)$ 時間と $O(hm/w + m\log(w)/w)$ 領域で解くビット並列アルゴリズムを与える. ここで, m と n はパターン木とテキスト木のサイズであり, h はテキスト木の深さ, w は計算機ワード長である.

第 5 章では, 正規表現の部分クラスに対するビット並列アルゴリズムに基づいたパターン照合ハードウェアの実装について考察する. 本章では, とくに, FPGA (field programmable gate array) と呼ばれる再構成可能ハードウェアを用いる. FPGA を用いたパターン照合ハードウェアは, 2001 年に Sidhu と Prasanna が Thompson の NFA に基づいた設計を与えて以来, 主に回路設計分野で研究されてきた. 一方で, その多くはパターンを回路構成時に静的に設定する手法であり, パターンを回路構成後に動的に設定できる手法はほとんどない. 2006 年に Baker らは, 決定性有限オートマトンとマイクロコントローラに基づいた正規表現照合ハードウェアを提案した. この手法は, パターンを動的に設定できる. 一方で, その照合性能はテキストの内容に依存する. そこで, 本章では, ビット並列アルゴリズムに基づいたパターン照合ハードウェアの基本設計を与える. この手法は, パターンを動的に設定できるだけでなく, 最悪時照合性能の理論的保証を与えている. また, ビット並列アルゴリズムを基本設計に適用することで, さまざまなパターン照合問題に拡張可能である. また, 拡張文字列に対するビット並列アルゴリズムである拡張 Shift-And 手法, および, 第 3 章で与えたネットワーク表現に対するビット並列アルゴリズムに基づいたパターン照合ハードウェアの評価実験の結果を示す. 提案手法は, 2004 年に Baker らが提案した文字列に対するパターン照合ハードウェアと, 2006 年に Yang らが提案した文字列に対するパターン照合ハードウェア, 2006 年に Baker らが提案した正規表現に対するパターン照合ハードウェアと同等のパターン照合性能を達成した.

最後に, 第 6 章で本論文の結論と今後の課題を述べる.

学位論文審査の要旨

主　査　教　授　有　村　博　紀
副　査　教　授　湊　　真　一
副　査　教　授　宮　永　喜　一
副　査　准教授　喜　田　拓　也

学位論文題名

Efficient Regular Expression Matching Algorithms and Their Implementation on Reconfigurable Hardware

(効率良い正規表現照合アルゴリズムとその再構成可能ハードウェア上の実装)

本論文では、理論情報科学における重要な問題の一つである正規表現照合アルゴリズムの開発と、並列ハードウェアを用いたその高速な実装法について研究している。正規表現照合問題は、長さ m の正規表現 (パターン)P と長さ n の長い文字列 (テキスト)T が与えられたとき、P の T 中の出現位置をすべて見つける問題であり、古典的かつ重要な計算問題として、古くから研究されてきた。一方 2000 年代に入って、大規模ストリーム処理のための並列ハードウェア上での高速実装が注目され、盛んに研究が行われている。

理論的観点からは、正規表現パターン照合問題において、素朴な $O(mn)$ 時間アルゴリズムの高速化が 1970 年代以来の未解決問題であった。これに対して、1992 年に Myers は、パターンを長さ k の断片に分割して計算する方式を用いて、初めて $O(mn/k)$ 時間でこの問題を解くアルゴリズムを与え、$O(mn)$ 時間の壁を破った。しかし、この方式は使用領域が大きく、実装が難しい。これに対して、2001 年に Navarro と Raffinot は、先行する Wu と Manber (1992) および Baeza-Yates と Gonnet (1992) らの文字列照合に関する結果を拡張し、ビット並列手法という技法を用いて、ビット幅が w のときに、拡張文字列パターンと呼ばれる直線状の正規表現の部分クラスに対して $O(mn/w)$ 時間の照合アルゴリズムを与えた。この方式は、パターンをビットマスクにコンパイルし、低レベルのブール演算と算術演算で照合を実行するため、実装上の効率が良く、FPGA や GPU などのハードウェアへの高速実装に適している。一方で、分岐や繰り返しを含むような一般的な正規表現への拡張は難しく、大規模ストリーム処理等への応用に限界があった。

そこで本論文では、分岐を含むような一般的な正規表現を扱えるようにビット並列手法を一般化し、さらにこの手法に基づいて FPGA と呼ばれる並列ハードウェア上の効率よい実装方式を確立することを目的として研究を行っている。

第 2 章では、正規表現照合問題を導入している。第 3 章では、ネットワーク表現と呼ばれる分岐をもつが、繰り返しを含まない正規表現の部分クラスを考察し、構文木の深さが d であるネットワーク表現の照合問題を $O(ndm/w)$ 時間と $O(dm/w)$ 領域で解く効率よいアルゴリズムを与えている。これは深さの小さなネットワーク表現に対しては、現在、最も良い計算量の結果である。さらに、これを一般の正規表現のクラスに拡張し、$O(ndm\log(m)/w)$ 時間と $O(dm\log(m)/w)$ 領域のアルゴリズムを与えている。

次に、第 4 章では、上記の技法を木パターン照合問題に適用し、XML データ照合の形式的モデルである多対一写像を許した無順序木パターン照合問題を考察している。この問題に対して、新しいビット並列手法を用いて、テキスト木の深さ h に対して、$O(nm\log(w)/w)$ 時間と $O(hm/w + m\log(w)/w)$ 領域で解くアルゴリズムを与えている。また、枝の伸長を許した問題の効率良い解法も与え、従来のアルゴリズムの計算量を大きく改善している。

第 5 章では、第 3 章の結果に基づいた大規模正規表現照合システムのハードウェア実装について議論する。初めに、提案の拡張されたビット並列アルゴリズムに基づいて、FPGA 上での照合システムのアーキテクチャを与え、次に拡張文字列パターンとネットワーク表現の両方に対して、具体的なハードウェアアルゴリズムの高速実装法を与えている。評価実験では、従来手法に比べて、提案手法は動的再構成性と、パターンの表現力、最悪時評価の計算時間の 3 点で優位なことが示されている。

本論文の成果は、次のようにまとめられる:

1. 理論情報科学における正規表現照合問題に対して、ビット並列手法に基づくアルゴリズムの系統的な設計に取り組み、種々の問題に対して効率よい照合アルゴリズムを与えた。これにより、正規表現および木パターン照合における新しい結果を示した。

2. ハードウェア設計において、再構成可能ハードウェアにおける新しい高速パターン照合アーキテクチャを提案し、さらに照合アルゴリズムの FPGA 上の実装を与え、性能評価を行うことで、提案のアーキテクチャの有効性を実証した。

これを要するに、著者は、大規模パターン照合のための照合アルゴリズムとハードウェア設計の両者について、新しい方法論を提案し、種々の問題に対して効率よい解法が構成可能であるという新知見を得たものであり、理論情報科学とデータ工学において貢献するところ大なるものがある。よって著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。