



Title	Faster Pattern Matching Algorithms for Massive Regular Expression Matching and Its Applications [an abstract of dissertation and a summary of dissertation review]
Author(s)	笹川, 裕人
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12183号
Issue Date	2016-03-24
Doc URL	https://hdl.handle.net/2115/61619
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Hirohito_Sasakawa_review.pdf, 審査の要旨



学位論文審査の要旨

博士の専攻分野の名称 博士 (情報科学)

氏名 笹川 裕人

審査担当者 主 査 教 授 有村 博紀

副 査 教 授 湊 真一

副 査 准教授 喜田 拓也

学位論文題名

Faster Pattern Matching Algorithms for Massive Regular Expression Matching and Its Applications

(大規模正規表現照合のための高速なパターン照合アルゴリズムとその応用)

本論文では、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の種々の分野における情報検索問題への適用について考察している。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。一方で、実世界での大規模データ処理と情報検索応用の広がりに伴い、大規模並列検索や、時空間系列データ検索、プライバシー検索等のこれまでのテキスト検索を超えた新しい正規表現検索問題への拡張が重要な課題となっている。そこで、本論文では、正規表現検索のこれらの実世界応用への適用を理論的および実証的に研究している。

この正規表現照合問題を解く古典的なアルゴリズムとして、決定性有限状態オートマトン (DFA) への変換に基づく手法が広く用いられている。これは、与えられた正規表現を、前処理で DFA に変換し、実行時に DFA をテキスト上で実行する手法である。この手法は、実行時間がパターンの長さと無関係に線形であり、極めて高速であるが、一方で、複雑な正規表現に対しては DFA の状態数が指数的に爆発する問題をもつ。そのため、本論文で扱うような一般の正規表現照合に基づく種々の応用に拡張することは難しい。

これに対して、本論文では、1970 年代に提案された正規表現の非決定性オートマトン (NFA) への変換に基づくパターン照合の枠組みを取り上げて研究している。この枠組みは、前処理において入力正規表現を等価な NFA に変換し、実行時に NFA の計算をテキスト上で模倣して照合を行う。この手法は、状態爆発を起こすことなく広範なパターン族を扱える一方で、NFA の効率良い実行が難しく、1980 年代まではあまり知られていなかった。しかし、1990 年代に We と Manber らおよび、Gonnet と Baeza-Yates らのグループが、計算機ハードウェアがもつビット演算と数値演算を用いることで NFA の模倣を効率よく実行するビット並列計算と呼ばれる技法を提案してから、さまざまな大規模データ処理アルゴリズムで広く使われるようになっていく。本論文では、この NFA に基づく正規表現手法の枠組みと、ビット並列計算に基づいて、情報科学の様々な分野における拡張された正規表現照合問題を高速に解くアルゴリズムを与えている。

第 3 章では、並列ハードウェアの一つである GPU (graphic processing unit) 上での正規表現照合の効率良いアルゴリズムを考察している。入力パターンクラスとして文字と、文字集合、有限と無限の繰り返しの演算子を連結した直線状の正規表現である拡張文字列パターンを考えて、基本戦略として、入力パターンから変換して得られた NFA を、小さなモジュールに分割して状態遷移を並列に

実行する。しかし、そのままではモジュール長を超える連続した長い空遷移列の存在のために、広範囲に渡る大量の遷移計算が発生し、並列実行効率が低下する問題がある。それに対し提案アルゴリズムは、NFA を 2 階層の小さなモジュールに分けた階層 NFA を用い、連続する空遷移の計算を上部モジュール内部のビット並列計算に局所化することで、GPU の演算コアを効率的に用い高速な照合を実現している。計算機実験では、提案手法は、2010 年に Lin らが提案した AC アルゴリズムを拡張した並列照合アルゴリズムに対して、アミノ酸配列データ上で約 150 倍の高速化を達成している。

第 4 章では、近年特に増加が著しい移動体の軌跡データを対象とした正規表現照合問題を考察している。本章では、はじめに移動体の座標の時系列である軌跡データに対して、 L_∞ 距離を用いた類似度を定義し、その類似度の下で軌跡正規表現照合問題を高速に解くアルゴリズムを与える。この問題の難しさは、離散的な文字の代わりに連続無限な実数座標をもつ多次元空間では、文字から状態遷移への対応付けが難しい点である。そのために、あらかじめ与えられたパターンからデータ空間の領域分割を求めておき、実行時に与えられた入力パターン点が部分領域に含まれるかを高速に判定するための「点位置決定データ構造」を用いて、軌跡データに対する正規表現照合問題を $O(nm \log \log(n) / \log(n))$ 時間と、 $O(m)$ 領域で解いている。ここで、 m と n は、それぞれ軌跡正規表現と軌跡テキストのサイズである。

最後に 5 章では、大規模データの利活用に関する重要な問題であるプライバシー保護を考慮した秘密正規表現照合を考察している。ここではデータ提供者とデータ分析者が、秘密情報としてそれぞれ正規表現 R とテキスト T を持ち、お互いの秘密情報を相手に漏らすことなく、 R の T における出現の有無のみをデータ提供者に対して報告する問題を考察して、 $O(nm^2)$ 時間と、 $O(nm^2)$ コミュニケーション、 $O(nm)$ ラウンドで解く秘密 NFA 評価プロトコルを与える。ここで、 m と n は入力正規表現とテキストのサイズである。基本的アイディアは、暗号文の加算が可能な加法準同型暗号を用い、ビット並列手法に基づく NFA の遷移計算を、加法準同型暗号上の演算で実装することである。計算機実験では、2009 年に提案された DFA を評価する Troncoso らの手法に対して、提案手法が正規表現のサイズに関して指数的に高速であることを確認している。

これを要するに、著者は、実世界の多様な情報検索問題における正規表現照合の高速な実現において、従来の技術が持つ問題点の解析に基づいて、種々の正規表現照合問題を考察し、これらの問題に対して効率良いパターン照合アルゴリズムの設計と実装が可能であるという新知見を得たものであり、情報科学における大規模データ処理の分野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。