HOKKAIDO UNIVERSITY

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_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学 位 論 文 内 容 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

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）

近年のネットワークインフラと, センサー技術の急速な発展により, 大規模な半構造ストリームデータが, 様々な分野で発生している. ネットワークのアクセスログや, GPS による移動物体の軌跡データ, 生物の遺伝子データなどは, その例である. このような大規模なストリームデータから所望の情報を取り出すことは重要な問題である. そのための基本的技術が正規表現照合である. 正規表現照合問題は, 入力テキスト T 中において, 与えられた正規表現 R のすべての出現位置を報告する問題である. 正規表現照合問題は, 理論計算機科学における基本的な問題であり, 多くの理論的研究がなされている. また, その豊富な表現力から, 広い応用に用いられている. 正規表現照合技術の応用としては, 例えば, ネットワーク不正侵入検出や,GPS データ解析, 患者の遺伝子情報を用いた疾病診断などがあげられる.

その一方で, 現在の正規表現照合技術はいくつかの問題を持っている. それらの問題点のうちで本論文では次の 3 つの問題に着目する. (1) 一つの入力データに対し, 多数のパターンの照合を行う要求が出現している. これは, 従来の CPU 上の高速化のみでは計算負荷が高く効率の良い実現は難しい.(2) 近年, アクセスログや, 株価データ,GPS データへの検索要求が高まっているが, これらの多次元数値時系列データに対して, 従来の正規表現照合手法をそのまま適用することは難しい.(3) 個人ゲノムにおける特定の病気に関連したパターン検索が重要性を増している. 個人情報保護の観点からデータ保有者と解析者の間で相互に情報が漏洩しないようなプライバシー保護が可能な正規表現照合が必要である.

そこで本論文では, 正規表現照合を実世界大規模データに適用する際に生じる上記の問題に取り組み, 最新のパターン照合技術を用いてこれらの問題を解決する. これにより, 実世界大規模データに適用可能な効率良い正規表現技術を開発する. 問題解決のための鍵となる技術として, 本研究では,1990 年代初頭に開発されたビット並列照合手法 (bit-parallel matching method) に基づき, 新たなパターン照合アルゴリズムを提案する. ビット並列手法は, 計算機レジスタ内の数値演算の並列性を利用して, 高速なアルゴリズムを設計する技法である. ビット並列照合手法の一つである SHIFT-AND 法は,1992 年に Wu と Manber, Beaza-Yates と Gonnet によって提案された.SHIFT-AND 法は, 入力正規表現を非決定性有限オートマトン (non-deterministic finite automata) に変換し,NFA の状態遷移表を表すビットマスク集合を構築する. 実行時には, オートマトンの状態集合をビット列で表現し, 前処理で構築したマスク配列との論理演算とシフト演算を用いて NFA の遷移計算を高速に模倣し照合問題を解く. また, 文字集合や, 限定繰返し, 非限定繰返し, ワイルドカードなどを含むより複雑な正規表現に対して, 減算などの算術演算を追加することで, 効率良い拡張が可能である. 本研究では, SHIFT-AND 法に基づく NFA 遷移計算を用いて, それぞれの問題に対して, 効率良いアルゴリズムを開発する.

はじめに, 第 3 章では, 複数の正規表現パターンに対する照合問題である大規模正規表現照合問

題を考察する. 本章では, 並列ハードウェアの一つである GPU (graphic processing unit) を用いた並列処理による高速化を考える. 入力パターンクラスとして, 正規表現の部分クラスである拡張文字列パターン (extended string pattern) を考え, そのパターンクラスに対する GPU 上の並列照合アルゴリズムを与える. 拡張文字列パターンとは, 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. 文字列のみからなるパターン照合と異なり, このクラスの正規表現においては, 連続する空遷移により, 広範囲に渡る大量の遷移計算が発生してしまい速度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 (hierarchical NFA) を用い, 連続する空遷移の計算をモジュール内部の計算に局所化したことで, 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た.

次に, 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える. 軌跡データは, 移動体の座標の時系列である多次元の数値ベクトルデータである. 本章では, はじめに軌跡データに対する $L\infty$ 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定式化する. 多次元空間における正規表現照合においては, 軌跡テキストの点に対し, 各パターン点との照合の判定を高速に実行することが困難である. そこで, 提案アルゴリズムでは,$L\infty$ 距離の下で, 軌跡テキスト上の点に対して, パターン点が構成するどの平面領域が照合するかを高速に判定するための点位置決定データ構造 (point location data structure) を与える. このデータ構造を用いた遷移表の高速な表引きにより, 提案アルゴリズムは, ビット並列手法に基づく効率良い NFA の遷移計算を実現する. 提案アルゴリズムは, 軌跡データに対する正規表現照合問題を $O(nmloglog(n)/log(n))$ 時間と, $O(m)$ 領域で解く. 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.

第 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データ提供者とデータ分析者それぞれのプライバシーを考慮した正規表現照合を考察する. 本章では, パーティ A とパーティ B が, 秘密情報としてそれぞれ正規表現 R とテキスト T 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パーティ A に対して報告する秘密正規表現照合問題を考える. はじめに, 秘密正規表現照合問題を $O(nm^2)$ 時間, $O(nm^2)$ 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. 提案プロトコルは, 暗号文の加算が可能な加法準同型暗号 (additively homomorphic encryption) 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へと拡張したことで, 効率良い照合を実現した. また本章では, 入力パターンに対してエラーを許容した照合問題である秘密近似文字列照合を高速に解くプロトコルも与える. 計算機実験では, 2009 年に提案された決定性有限オートマトン (DFA, deterministic finite automata) 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た. 最後に, 第 6 章で本論文の結論と, 今後の課題を述べる.