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　元 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００００８７

研究課題名（和文）形式言語理論を駆使したウェブ基盤技術の検証

研究課題名（英文）Verification of Web technologies based on the theory of formal languages

研究代表者
南出　靖彦（Minamide, Yasuhiko）

東京工業大学・情報理工学院・教授

研究者番号：５０２５２５３１

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：ウェブ基盤技術の検証問題から出発し，様々な計算モデルに基づく検証技術の研究を行なった．主な成果としては，バックトラックに基づく正規表現マッチングの実行時間のオーダを決定する解析アルゴリズムの開発があげられる．正規表現からマッチングの計算過程を表現する先読み付き木トランスデューサを構築し，その増加率のオーダを決定している．増加率の解析は，AhoとUllmanによる先読みの無い木トランスデューサに対する解析手法を，先読み付きに拡張することで実現している．既存のPHP プログラムで使用されている正規表現を対象に実験を行い，オーダが2乗や３乗となる正規表現の検出に成功した．

研究成果の学術的意義や社会的意義
本研究で開発した正規表現マッチングの時間計算量解析はウェブソフトウェアのReDoS脆弱性の検出に直接応用可能なものであり，ウェブソフトウェアの信頼性の向上に貢献する．また，本研究で展開した形式言語理論の研究は，ウェブの基盤技術にとどまらずより一般のソフトウェア検証に応用できる．特に，正規表現マッチングの時間計算量解析の理論とストリーミングトランスデューサの合成に関する研究は，今後のソフトウェア検証研究の進展の基盤となるものである．

研究成果の概要（英文）：Starting from verification problems related to the foundation of Web technologies, we have investigated the theory and practice of verification based on various computation models. Our main contribution is that we developed an analysis that determines the order of execution time for regular expression matching based on backtracking. The analysis first constructs a tree transducer with lookahead simulating search of regular expression matching and then determines the growth rate of the tree transducer. The growth rate was determined by extending the analysis for tree transducers without lookahead by Aho and Ullman. We conducted experiments on regular expressions obtained from publicly available PHP programs and successfully found regular expressions with quadratic and cubic orders.

研究分野：ソフトウェア検証

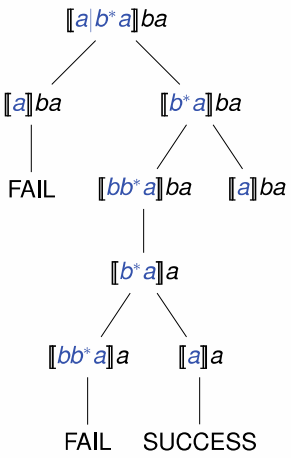
キーワード：ソフトウェア検証　形式言語理論　ウェブ　正規表現　プッシュダウンオートマトン　構文解析

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

現在のウェブ環境は，ウェブサーバ，ブラウザ，ブラウザ上で動作する JavaScript など非常に複雑な基盤の上に成り立っている．ウェブブラウザなどの基盤となるプログラムやその上で動作するウェブプログラムの誤りは，クロスサイトスクリプティングなどの脆弱性の原因となり，情報漏洩などの深刻な問題を起こし，ウェブソフトウェアの信頼性を損なう重大な問題となっている．このような問題を解決するために，形式言語理論に基づく検証技術が適用され，クロスサイトスクリプティング脆弱性の検出等において実用的にも有用な検査器が開発されるようになってきていた．

ウェブソフトウェアにおける新たな形態の脆弱性として，正規表現マッチングに起因する DoS 脆弱性(ReDoS 脆弱性)が問題となっていた．プログラミング言語における正規表現マッチングのほとんどの実装はバックトラックに基づいていおり，右の図のように成功するまで探索が行われる．そのため，最悪の場合，指数関数時間の計算量を持ち，DoS 脆弱性の原因となる．上記の問題を検出するため，バックトラックに基づく正規表現マッチングの時間計算量に関する研究を 2013 年から始めていた．先行研究において，木トランスデューサ（出力付き木オートマトン）の理論を適用し，正規表現マッチングの時間計算量が O(n)であるかを判定するアルゴリズムを開発した．しかし，実際のウェブプログラムにおける ReDoS 脆弱性の検査に適用するには，依然，解決すべき問題が多かった．先行研究で行ったアルゴリズムの実装は非効率的であり，実際のウェブプログラムから得られた複雑な正規表現に対して停止しない場合があった．また，脆弱性の原因となる入力例の提示ができておらず，判定結果の理解が非常に難しいという問題があった．



図：バックトラックによる正規表現マッチング

２．研究の目的

クロスサイトスクリプティング脆弱性の検出等の限定された問題に適用されてきた形式言語理論に基づく検証技術を，ウェブ環境を支える様々なソフトウェアに対して適用し，その信頼性の向上を目指す．同時に，ソフトウェア検証の基盤となる計算モデルの理論，形式言語理論の研究を行い，ウェブソフトウェア検証の基盤となる理論を進展させ，その応用を進める．また，DoS 脆弱性などより広いクラスの脆弱性の検査技術についても研究を進める．特に，正規表現マッチングに起因する ReDoS 脆弱性について，木トランスデューサ理論に基づく検査技術を確立する．

１）ReDoS 脆弱性の木トランスデューサに基づく検査技術を確立する．先行研究で開発した正規表現マッチングの時間計算量が線形であるかのアルゴリズムを詳細に分析し，実際のウェブプログラムから得られる正規表現に関して，十分に効率的な実装を得る．また，検査結果の有用性を高めるために，$w_1w_2^nw_3$ のような形で非線形性を示す具体例を提示する手法を開発する．さらに，非線形性による脆弱性の危険度を評価するために，$O(n^2)$, $O(n^3)$などの階層のための判定アルゴリズムを開発する．

２）ウェブプログラム検証の基盤として，高度な計算モデルを用いた文字列操作プログラムの検証技術を確立する．これまでの研究代表者らの研究では，文字列を操作するプログラムをトランスデューサを用いて表現してきた．しかし，文字列の置換などの操作は正確に表現できないため，保守的な近似を適用して表現する必要があり，偽陽性の原因となっていた．この問題を解決し，より精度の高い検証を可能とするために，Alur らによって導入されたストリーミングトランスデューサなどのより表現力の高い計算モデルを用いた検証技術を確立する．

３．研究の方法

ウェブの基盤となる技術・プログラムの検証に形式言語理論を適用し，ウェブソフトウェアの信頼性を高める研究を行う．　現在のウェブ環境は，ウェブサーバ，ブラウザ，サーバサイド・プログラムなど様々な構成要素から成立っており，構成要素によって要求される安全性などの性質は異なるが，HTML などを文字列として受け渡ししているという点では共通しており，形式言語理論による検証手法が有効であると考えられる．

正規表現マッチングに関する研究においては，反例の提示を早期に実現し，ReDoS 脆弱性に関する実験を実際のサーバサイドプログラムに対し行い，その結果に基づき検査手法を改良し，実用的な検査ツールの実現に繋げる．同時に，検証の基盤となるプッシュダウン・システムの拡張及び文字列，木，グラフ上のトランスデューサの理論の研究も進める．

４．研究成果

ウェブ基盤技術の検証のための形式言語理論および様々な基盤技術に関する具体的検証技術について研究し，以下の成果を得た．

１）正規表現マッチングに起因する ReDoS 脆弱性を検出する研究を行った．具体的にはバックトラックに基づく正規表現マッチングの実行時間のオーダを決定する解析アルゴリズムを開発した．正規表現からマッチングの計算過程を表現する先読み付きトップダウン木トランスデューサを構築し，その増加率のオーダを決定している．増加率の解析は，Aho と Ullman による先読みの無いトップダウン木トランスデューサに対する解析手法を，先読み付きに拡張することで実現している．この解析アルゴリズムを実装し，既存の PHP プログラムで使用されている正規表現を対象に実験を行い，オーダが 2 乗や３乗となる正規表現の検出に成功した．また，そのオーダの振る舞いを示す入力パターンを生成でき，脆弱性の原因の理解を支援している．さらに，この計算量解析に関して，効率的な判定器を実装する研究を行った．Weber と Seidl によるオートマトンの曖昧度の判定アルゴリズムの手法と本研究で開発してきた木トランスデューサに基づく手法を組み合わせることで効率的な判例アルゴリズムを構築した．これまでの実装でタイムアウトになっていた正規表現についても実用的な時間内での判定が可能となった．

２）時間の概念を取り入れたプッシュダウン・オートマトンの研究を行った．時間プッシュダウン・オートマトンにおける到達可能性問題が決定可能であることは，2012 年に Abdulla らによって示されていたが，その証明は非常に複雑であり，表現力も不自然に弱い体系となっていた．それに対して，本研究は小数部検査制約を導入し，さらに体系を一般化することで，Synchronized Recursive Timed Automata という体系を提案し，小数部検査制約により言語の表現力が広がることを証明した．また、到達可能性問題の決定可能性について，backward-forward simulation を用いて見通しの良い証明を与え，既存の結果よりも強い計算状況に対する到達可能性が決定可能であることを示した．

３）インデックスされた重み領域を持つ重み付きプッシュダウンオートマトンの到達可能性判定に関する研究を行った．この問題の判定アルゴリズムは，先行研究において HTML5 構文解析仕様に対するテストの自動生成に応用されている．この判定アルゴリズムの基礎となる理論を整理し，また，この拡張された重み付きプッシュダウンオートマトンの枠組みで，スタックを書き換え可能にしたプッシュダウンオートマトンおよび Well-structured プッシュダウンオートマトンの到達可能性判定アルゴリズムを説明できることを示した．

４）ウェブの基盤技術の検証に有効であると考えられる Streaming String Transducer に関して，合成法の研究を行った．Alur らによる先行研究によって，Streaming transducer が合成に関して閉じていることが示されていたが，詳細な構成法を示されておらず，また，最初に与えらえた構成法に誤りがあることが知られていた．本研究では，この合成の構成法を詳細に定義し，厳密な証明を与えた．さらに，Streaming Transducer のコピーに関する制限を緩和した Bounded Streaming String Transducer を導入し，合成の具体的構成を与えた．また，証明支援系 Isabelle/HOL により Bounded Streaming Transducer の形式化を行い，合成の構成の正しさを証明した．

５）先読み付き正規表現に，形式言語理論的な解釈を与え，その理論を構築する研究を行なった．文字列上の関係に連接，クリーネスター等の演算を導入した先読み付き正規言語がクリーネ代数を構成すること示した．これにより，クリーネ代数の枠組みを用いて，先読み付き正規表現の意味が先読み付き言語として自然に解釈できるようになった．さらに，正規表現における微分が，先読み付き正規表現に自然に拡張できることを示した．この微分を用い，先読み付き正規表現からオートマトンへの変換を与え，このようなオートマトンへの変換の状態数に関する上界が二重指数であることを示した．

６）ウェブブラウザ上で実装されているクロスサイトスクリプティング保護機構である XSS Auditor の有効性を検証する研究を行った．XSS Auditor で実行される検査を二つのトランスデューサの包含判定としてモデル化した．トランスデューサの包含判定は一般には決定不能問題であるため，保守的な検査アルゴリズムを実装した．非常に限定された状況においては，XSS Auditor の保護機能が有効であることが検証できた．

５．主な発表論文等

〔雑誌論文〕（計５件）

① Yuya Uezato, Yasuhiko Minamide, Synchronized Recursive Timed Automata, International Conference on Logic for Programming, Artificial Intelligence, and Reasoning, LNCS 9452, 249-265,2015.査読有.
DOI:10.1007/978-3-662-48899-7_18

② Yasuhiko Minamide, Weighted Pushdown Systems with Indexed Weight Domains, Logical Methods in Computer Science, Vol. 12(2:9), 2016, pp. 1-27. 査読有
DOI: 10.2168/LMCS-12(2:9)2016

③ Yuya Uezato and Yasuhiko Minamide, Monoid-based Approach to the Inclusion Problem on Superdeterministic Pushdown Automata, DLT 2016: Developments in Language Theory, pp. 393-405, 2016. 査読有.
DOI:10.1007/978-3-662-53132-7_32

④ Yuya Uezato and Yasuhiko Minamide, Configuration Reachability Analysis of Synchronized Recursive Timed Automata, コンピュータソフトウェア, Vol. 35, No. 1, p. 140-168, 2018. 査読有.
DOI:10.11309/jssst.35.1_140

⑤ Takayuki Miyazaki, Yasuhiko Minamide, Derivatives of Regular Expressions with Lookahead, Journal of Information Processing, 2019. 査読有.

〔学会発表〕（計７件）

① 上里 友弥, 南出 靖彦, 更新可能時間オートマトンの新たな拡張について, 日本ソフトウェア科学会第３２回大会,2015 年 9 月

② 中川みなみ, 南出 靖彦, バックトラックによる正規表現マッチングの時間計算量解析, 第 107 回プログラミング研究会, 2016 年 1 月.

③ 赤間 仁志, 南出 靖彦, Streaming String Transducer の合成の形式的証明（ポスター）, プログラミングおよびプログラミング言語ワークショップ,2018 年 3 月.

④ 高橋 和也, 南出 靖彦, 言語の包含判定に基づくサニタイズ文脈の自動決定（ポスター）, プログラミングおよびプログラミング言語ワークショップ,2018 年 3 月.

⑤ 宮嵜 貴之, 南出 靖彦, 先読み付き正規表現の微分, 情報処理学会プログラミング研究会第 121 回プログラミング研究発表会, 2018 年 11 月.

⑥ 宮嵜 貴之, 南出 靖彦, 先読み付き正規表現と解析表現の微分（ポスター）, プログラミングおよびプログラミング言語ワークショップ,2019 年 3 月.

⑦ 高橋 和也, 南出 靖彦, バックトラックによる正規表現マッチングの計算量判定の実装ポスター）, プログラミングおよびプログラミング言語ワークショップ,2019 年 3 月.

〔その他〕
ホームページ：http://sv.c.titech.ac.jp/

６．研究組織

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。