平成24年度 情報工学コース卒業研究報告要旨

坂部研究室 研究室	氏　　名	林　篤　弘
卒業研究題目	「.」メタ文字を含む正規表現の同値性判定に関する研究	

正規表現は，文字列のパターンを表現する表記法であり，各種プログラミング言語，コンパイラ，文字列の検索や置換などに利用されている．形式言語理論において，アルファベットΣ上の正規表現は，連接（·）や選言（+）、クリーネ閉包（*）を用いて表される．プログラミング言語では，利便性を高めるため，メタ文字を含む拡張された正規表現が利用されている．区別のため，形式言語理論における正規表現は形式正規表現，メタ文字を含む拡張された正規表現を拡張正規表現，これらをまとめたものを正規表現と呼ぶ．正規表現と言語のパターンマッチ処理の効率化などのため，正規表現の簡略化は重要である．簡略化のために，異なる形の正規表現の同値性を判定することも重要である．異なる形の正規表現が同一の言語を表しているかを判定する基本的な方法として，正規表現を決定性有限オートマトン（DFA）に変換し比較するという手法がある．これに対し，Antimorv, Mosses らにより，アルファベットΣ上の2つの形式正規表現を線形正規表現に書き換え，先頭の文字以降の形式正規表現を状態に加え，状態を増やしつつ等価性を判定する操作を加えることで，異なる形式正規表現の同値性を判定する手法が提案された．線形正規表現とは，$a_i \in \Sigma$，$\alpha_i \in$ としたとき，$a_1\alpha_1 + \cdots + a_n\alpha_n$ で表される正規表現である．そして Almeida，Moreira らによる実装，実験により，この手法は DFA に変換する手法より高速に同値性を判定できることが示された．しかし，この手法は形式正規表現しか扱っておらず，拡張正規表現には対応していない．

本研究では、利便性を高めるために，このアルゴリズムを各種言語に適用することを考える．よく利用されている正規表現では，[afgoz] で [] 内の文字の集合，?で直前の一文字が0または1回,「.」で任意の一文字など，特定の文字列のパターンを表すためのメタ表現が利用されている．これらは，[afgoz] は a+f+g+o+z，などと書き換えることによりアルゴリズムが適用できるが,「.」の場合はその言語に含まれる文字全体の集合となり，その言語で扱える文字数分，状態数が多くなるという問題がある．ASCII では 7bit で 128 文字か 8bit で 256 文字ほどであるが，JIS 第1水準では約 7000 文字分，超漢字では約 19 万文字分状態が増えるため，処理が膨大になる．したがって,「.」メタ文字を扱う拡張正規表現では，Antimorv，Mosses らの手法をそのまま用いることは現実的ではない．

そこで,「.」メタ文字を含む正規表現の同値性を判定できるよう，Antimorv，Mosses らが提案した手法を拡張する．そのために，アルゴリズムに「.」メタ文字を扱う処理を追加し，その文字を含む正規表現の同値性を判定するアルゴリズムを提案する．そして，提案したアルゴリズムを Standard ML of New Jersey で実装し，適当な正規表現を用いて検証を行う．

検証の結果,「.」メタ文字を含む正規表現の同値性も判定できることが示された．今後の課題としては，特定の文字列の n 回の繰り返しなど，正規言語ではない正規表現にも対応するアルゴリズムへの拡張，アルゴリズムの停止性の証明などがあげられる．