A-011

# 正規表現の接頭辞及び接尾辞を用いた NFA 構成法の実験的評価
# On the Experimental Evaluation of an NFA Construction based on Prefixes and Suffixes of Regular Expressions

南波 龍一† 宮本 純平† 山本 博章†
Ryuichi Nanba Junpei Miyamoto Hiroaki Yamamoto

## 1. はじめに

正規表現に対するパターン照合問題は，情報セキュリティ，バイオインフォマティクス，情報検索などコンピュータサイエンスの多くの分野で応用されている．一般に，正規表現のパターン照合は，正規表現を有限オートマトンに変換して行う．そのため，正規表現から効率的な有限オートマトンを生成することは重要であり，その手法は古くから研究されている．代表的なものは，Thompson オートマトンである．さらに，状態数の少ない有限オートマトンとして，位置オートマトン，等価オートマトン，フォローオートマトンなどが知られている．

山本[1]は，等価オートマトンの発展として，接頭辞オートマトン及び接尾辞オートマトンという新たなオートマトンの作成法を示し，接尾辞オートマトンが等価オートマトンと同じになることを示した．さらに，接頭辞と接尾辞を組み合わせたオートマトンも示し，効率的な構成アルゴリズムを与えた．また，Ko と Han[2]は，接頭辞及び接尾辞と同等の概念として左不変及び右不変の同値関係を導入し，左不変を用いて作成した NFA は右不変を用いて作成した NFA よりも，より決定性に近くなるという結果を示している．

これらの結果を踏まえ，本論文は，山本が開発した新たな構成法について，実験的に評価するものである．

## 2. 準備

正規表現の定義を与える．

【定義】アルファベット Σ 上の正規表現は以下のように定義される．

1. ∅，ε，a は正規表現である．
2. $r_1r_2$ を正規表現としたとき，$(r_1+r_2)$, $(r_1r_2)$, $r_1$*は正規表現である．ここで，各正規表現は，$R_1$, $R_2$ を $r_1$, $r_2$ が表す言語（語の集合）としたとき，それぞれ，言語 $R_1R_2$, $R_1R_2$, $R_1$*を表す．

正規表現の各演算子の優先順位は，高い方から閉包，連接，和の順とし，不要なカッコを省いて記述する．また，L(r)によって，正規表現 r が表す言語とする．正規表現の長さは m で表し，これは正規表現に現れるアルファベット記号と演算子の数と定義する．

## 3. Thompson オートマトン

正規表現から得られる NFA で最初に示されたのが Thompson オートマトン(T-NFA)である．T-NFA は，正規表現の定義にしたがって再帰的に構成される．構成法の概略を図 1 に示す．

†信州大学 Shinshu University

## 4. 接頭辞及び接尾辞を用いたオートマトン

我々の新たな構成法は，ラベル付き T-NFA を構成することから始まる．ラベル付き T-NFA とは，T-NFA の各状態に，正規表現をラベル付けしたものである．各状態のラベルは，初期状態からその状態に到達可能な語の集合を表す正規表現（接頭辞表現），および，その状態から受理状態に到達可能な語の集合を表す正規表現（接尾辞表現）の２つからなる．これらのラベルによって，状態の等価性を判定し，接頭辞オートマトン(PreEA)，接尾辞オートマトン(SufEA)を構成する．

### 4.1 ラベル付き Thompson オートマトン

ラベル付き T-NFA M=(Q,Σ,δ,$p_0$,$q_0$,LP,LS)を定義する．これは，T-NFA の各状態に接頭辞，接尾辞を表すラベルつけたものである．定義中の，Q は状態の集合，Σはアルファベット，δは状態遷移関数，$p_0$ は初期状態，$q_0$ は受理状態である．また，LP 及び LS は，M の状態から正規表現への関数で，それぞれ，接頭辞ラベル，接尾辞ラベルを表し，その値は下記の接頭辞表現及び接尾辞表現のラベル付け方法によって定義される．

【接頭辞表現ラベル付け手法】

1. 基本ラベル付け：初期状態 $p_0$，受理状態 $q_0$ に対し，
   (ア) もし r=∅ならば，$LP(p_0)=\varepsilon$ かつ $LP(q_0)=\emptyset$
   (イ) もし $r=\varepsilon$ ならば，$LP(p_0)=\varepsilon$ かつ $LP(q_0)=\varepsilon$
   (ウ) もし r=a ならば，$LP(p_0)=\varepsilon$ かつ $LP(q_0)$=a
2. $r_1$, $r_2$ を正規表現とし，$M_1$ 及び $M_2$ を $r_1$,$r_2$ に対するラベル付き T-NFA とし，その接頭辞のラベル付け関数を，それぞれ $LP_1$, $LP_2$ とする．そのとき，M の LP を以下のように定義する．
   (ア) $r=r_1+r_2$ のとき．$LP(p_0)=\varepsilon$，$LP(q_0)=(LP_1(q_1)+LP_2(q_2))$．その他の状態 q に対しては，もし q が $M_1$ の状態ならば，$LP(q)=LP_1(q)$，$M_2$ の状態ならば，$LP(q)=LP_2(q)$とする．
   (イ) $r=r_1r_2$のとき．すべての状態 q に対し，
   ① q が $M_1$の状態ならば，$LP(q)=LP_1(q)$
   ② q が $M_2$ の状態ならば，$LP(q)=(LP_1(q_1)LP_2(q))$とする．
   (ウ) $r=r_1$*のとき．$LP(p_0)=\varepsilon$，$LP(q_0)=(LP_1(q_1))$*とし，その他の状態 q に対しては，$LP(q)=(LP_1(q_1))$*$LP_1(q)$とする．
   (エ)

【接尾辞表現ラベル付け手法】

1. 基本ラベル付け：初期状態 $p_0$，受理状態 $q_0$ に対し，
   (ア) もし r=∅ならば，$LS(p_0)=\emptyset$かつ $LS(q_0)=\varepsilon$
   (イ) もし $r=\varepsilon$ ならば，$LS(p_0)=\varepsilon$ かつ $LS(q_0)=\varepsilon$



Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

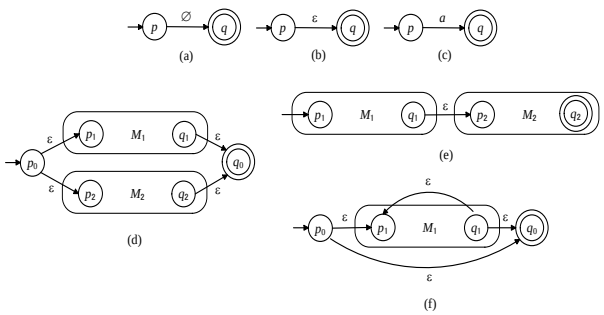

図 1　Thompson オートマトンの構成法

(ウ) もし r=a ならば，$LS(p_0)$= a かつ $LS(q_0)$= ε

2. $r_1$, $r_2$ を正規表現とし，$M_1$ 及び $M_2$ を $r_1$,$r_2$ に対するラベル付き T-NFA とし，その接尾辞のラベル付け関数を，それぞれ $LS_1$, $LS_2$ とする．そのとき，M の LS を以下のように定義する．
   (ア) $r=r_1+r_2$ のとき．$LS(p_0)=(LS_1(p_1) + LS_2(p_2))$，$LS(q_0) = \varepsilon$．その他の状態 q に対しては，もし q が $M_1$ の状態ならば，$LS(q)=LS_1(q)$，$M_2$ の状態ならば，$LS(q)=LS_2(q)$とする．
   (イ) $r=r_1r_2$のとき．すべての状態 q に対し，
      ① q が $M_1$ の状態ならば，$LS(q)=(LS_1(q)$ $LS_2(p_2))$,
      ② q が $M_2$ の状態ならば，$LS(q) = LP_2(q)$.
   (ウ) $r=r_1$*のとき．$LS(p_0)= (LS_1(p_1))$*，$LS(q_0)=$ ε とし，その他の状態 q に対しては，$LS(q) = LS_1(q)$ $(LS_1(p_1))$*とする．

### 4.2 接頭辞オートマトン及び接尾辞オートマトン

今，M=(Q,Σ,δ,$p_0$,$q_0$,LP,LS)をラベル付き T-NFA とする．Q'を，M の初期状態とすべての記号状態の集合とする．ここで，記号状態とは，T-NFA の状態で，その状態に入る遷移がすべてアルファベット記号によるものである．その時，Q'に，LP,LS に基づいた同値関係を定義する．

【定義】任意の状態 p,q∈Q'に対し，$p \equiv_{pre} q$ ($p \equiv_{suf} q$)とは，LP(p)=LP(q) (LS(p)=LS(q))であるときである．

上記の等価関係を用いて，接頭辞オートマトン及び接尾辞オートマトンを定義する．

【接頭辞オートマトン】$PE=(Q_P,\Sigma,\delta_P,[p_0],F_P)$を以下のように定義する．

- $Q_P$={[q]｜[q]は，$\equiv_{pre}$による同値類で q を含むもの}
- $[p_0]$：初期状態で，$p_0$ は M の初期状態
- $[q]\in\delta_P([p],a)$なる必要十分条件は[p]の中に記号 a で[q]の状態に到達できる状態があるときである．
- $[p]\in F_P$ となる必要十分条件は[p]の中に M の受理状態へ到達可能な状態が存在するときである．

【接尾辞オートマトン】$SE=(Q_S,\Sigma,\delta_S,[p0],F_S)$を以下のように定義する．

- $Q_S$={[q]｜[q]は，$\equiv_{suf}$による同値類で q を含むもの}
- $[p_0]$：初期状態で，$p_0$ は M の初期状態
- $[q]\in\delta_S([p],a)$なる必要十分条件は[p]の中に記号 a で[q]の状態に到達できる状態があるときである．
- [p]∈F となる必要十分条件は[p]の中に M の受理状態へ到達可能な状態が存在するときである．

### 4.3 接頭辞と接尾辞の統合オートマトン

接頭辞による状態の統合と接尾辞による状態の統合の両方を行ってできるオートマトンである．方法として，2 通りある．一つは，まず，接頭辞オートマトンを求め，そのあと，接頭辞オートマトンに対し，接尾辞による状態の統合を行う方法，もう一つは，接尾辞オートマトンを求め，そのあとに，接頭辞による統合を行う方法である．前者で得られるオートマトンを，接頭辞・接尾辞オートマトン(PreSufEA)と呼び，後者を，接尾辞・接頭辞オートマトン（SufPreEA）と呼ぶ．

## 5. 実験

表 1　記号数 4，長さ 20 の正規表現に対する各オートマトンのサイズの比較

| | 状態数 | 状態遷移数 | DFA 数 |
|---|---|---|---|
| PreEA | 14.28 | 18.48 | 19 |
| SufEA | 11.72 | 15.28 | 12 |
| PreSufEA | 11.46 | 14.9 | 20 |
| SufPreEA | 11.4 | 14.78 | 16 |

表 2　記号数 10，長さ 40 の正規表現に対する各オートマトンのサイズの比較

| | 状態数 | 状態遷移数 | DFA 数 |
|---|---|---|---|
| PreEA | 33.3 | 38.48 | 37 |
| SufEA | 28.88 | 33.58 | 25 |
| PreSufEA | 28.54 | 33.34 | 37 |
| SufPreEA | 28.52 | 33.2 | 28 |

提案法は，JAVA を用いて実装した．また，正規表現は FAdo[3]を用いて生成した．実際，FAdo によりランダムに正規表現を 50 個生成し，各表現に対し，接頭辞オートマトン(PreEA)，接尾辞オートマトン(SufEA)，接頭辞・接尾辞オートマトン(PreSufEA)，接尾辞・接頭辞オートマトン(SufPreEA)を作成し，各オートマトンの状態数及び状態遷移数，DFA 数（50 個中，最終形が DFA になった数）を調べた．各オートマトンの状態数，状態遷移数は 50 個の平均値を取った．表 1 は，アルファベットのサイズが 4，長さ 20 の正規表現に対する結果，表 2 はアルファベットサイズ 10，長さ 40 の正規表現に対する結果である．今回の結果から，PreEA が，状態数は多いが DFA になる割合が高いといえる．また，PreSufEA と SufPreEA を比較した場合，状態数，遷移数はほとんど変わらないが，PreSufEA のほうが DFA になる率が高いため，こちらのほうが良いと思われる．

### 参考文献

[1]H. Yamamoto, “A New Finite Automaton Construction for Regular Expressions”,Proc of NCMA2014, pp.249-264 (2014).
[2]S.Ko, Y.Han, “Lestis Better than Right for Reducing Nondeterminism of NFAs”,Proc. of CIAA 2014, LNCS 8587, pp.238-251 (2014).
[3]A. Almeida, M. Almeida, J.Alves, N.Moreira, R.Reis FAdo and GUItar: Tools for Automata Manipulation and Visualization, Proc. opf CIAA 2009, LNCS5642, pp.65-74(2009).



Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.