正規表現の特徴量を用いた同一判定の効率化

呉　晟董[†1]　成　　凱[†1]

概要：正規表現は文字列のパターンを正確に表現する言語として文字列処理に広く応用されている。しかし、同じパターンを記述するには複数の正規表現が正解になる可能性があり、正規表現の理解や最適な正規表現を求めるために正規表現の同一判定が必要である。従来手法として、正規表現を有限オートマトンに変換し、オートマトンの同一性から正規表現の同一性を判定することが一般的である。しかし、正規表現によって複雑な有限オートマトンになり、同一判定の効率が悪い問題が指摘されている。本研究では、正規表現の特徴量（要素数、文字列の長さ）やマッチングテストを利用して、同一判定の効率化を図る。正規表現の特徴量がオートマトンの状態数や記号列のサイズに依存せず、非同一判定の効率化に大きく寄与できる。評価実験により、提案手法の有効性が検証できた。

キーワード：正規表現，正規言語，有限オートマトン，テキスト処理，同一判定

1. はじめに

正規表現は、文字列のパターンを記述する言語の一つであり、文字列の検索や置換、抽出などを行う際の対象の指定などのために用いられる[10][11][12][13]。正規表現は有限オートマトンによって受理できる言語と対応することができる[9][14]。一方、正規表現の自由度が非常に高く同じパターンを表現するには複数の正規表現を書くことができるため、初心者にとっては、ある正規表現が正解と一致するかを判断することが難しい。正規表現の同一判定に関して様々な先行研究が多く存在するが、処理時間がΣ（アルファベット）の大きさ及び有限オートマトンの状態数に依存しているため、効率が悪い問題が指摘されている[3][8]。

本研究では、正規表現の特徴量（要素数、文字列の長さ）やマッチングテストを利用して同一判定の効率化を提案する。正規表現の特徴量がオートマトンの状態数や記号列のサイズに依存せず、非同一判定の効率化に寄与できる。提案手法を評価するために、任意の正規表現に対応する言語の最小長さ、最大長さ、要素数を求めるアルゴリズムを開発し、特徴量、マッチングテストによる同一判定を Perl で実装し、評価実験を行う。

2. 関連研究

正規表現の同一判定は計算理論の基本的問題として研究されている。まず、正規表現と有限オートマトンの関係に基づき、同一判定を行う手法が最もよく利用される。中では、最小確定有限オートマトン（DFA）の唯一性により、正規表現に対応する最小 DFA を求め、同一判定を行う方法が提案された[2]。しかし、k=|Σ|、状態数 n のオートマトンの最小化に最悪計算量が$O(kn\ \log n)$であり、計算コストが高いと知られている。これに対して、Hopcroft と Karp らは状態の同値性に基づき、最小化を必要としない同一判定アルゴリズム（HK アルゴリズム）が提案された[3]。HK アルゴリズムは同じΣ上で状態数がそれぞれ n と m の DFA の同一判定に必要な計算量が最悪$O(k(n+m))$であり、Σの大きさと状態数に依存している。

オートマトンを使わない同一判定アルゴリズムについて研究されていた。M.Almeid ら[8]は正規表現の代数的性質を利用した同一判定を行う手法を提案した。この手法は Rewrite 公理[4] [5][6]に基づいておいるが、Σの大きさに依存するため、性能面では上記の有限オートマトン法と変わらない。

3. 正規表現の同一判定

本節は、正規表現とその同一判定の基本的事項について述べる。

3.1 正規表現とその同一性

正規表現は，ある記号集合（アルファベット）上で定義される文字列のパターンを定義する言語である。文字列のパターンマッチングによく使われる。アルファベットΣ上の正規表現はメタ文字と通常文字に分けている。表 1 はよく使われるメタ文字を列挙している。

表 1　正規表現のメタ文字

記号	説明	記号	説明
.	任意 1 文字	*	0 回以上繰り返し
[ ]	[]中任意 1 文字	+	1 回以上繰り返し
[^]	[]外任意 1 文字	?	0 回か 1 回出現
¥d	数字 1 文字	{n}	n 回繰り返し
¥w	英数 1 文字	{n,m}	n 回以上 m 回以下繰り返し

正規表現で表すことができる言語は正則言語とも呼ばれ、有限オートマトン（DFA、NFA）によって受理できる言

1 九州産業大学大学院情報科学研究科
Graduate School of Information Science.
Kyushu Sangyo University

語と等しい。正規表現 α に対応する正規言語は$L(\alpha)$と表す。

定義：$L(\alpha) = L(\beta)$であれば、正規表現 α と β が同一正規表現と呼ぶ。

α と β が形式上で違っても同じ正規言語を表すものであれば、お互いに同一である。例えば$[ab]\{3\}$と$(a|b)\{3\}$は同じ言語$L = \{aaa, aab, aba, abb, bbb, bba, bab, baa\}$を表しているため、同一の正規表現である。

3.2 オートマトンに基づく同一判定

正規表現 α と β に対する正則言語$L(\alpha)$と$L(\beta)$を受理する有限オートマトンの同一性によって、α と β の同一性を判定する手法について述べる。まず、有限オートマトンの状態の同値性を説明する。

ある有限オートマトンにおいて、状態pとqが全ての入力列wに対して、$\hat{\delta}(p,w)$が受理状態で$\hat{\delta}(q,w)$も受理状態あるときにpとqは**同値**である。ここの$\hat{\delta}(p,w)$は状態 p で入力列 w を受けてその遷移先の状態を求める状態遷移関数である。直観的に言うと、二つの状態pとqが同値であるとは、これらのうちの一方を初期状態として与えられた入力列を読み進むとき受理状態に到達するかどうかを調べるだけでは、両者を区別することは不可能である。二つの状態が同値でないとき、その二つの状態は区別可能であるという。すなわち、状態pが状態qから区別可能であるのは$\hat{\delta}(p,w)$と$\hat{\delta}(q,w)$のうち一方は受理状態であるが、もう一方は受理状態ではない、となるような文字列wが少なくとも一つ存在することである。

同値な状態を全て見つけるには、区別可能な状態の対を可能な限り見つけ出せば良い。状態の同値を調べるには穴埋めアルゴリズムと呼ばれるアルゴリズムが提案されている[3][9]。これは DFA $A = (Q, \Sigma, \delta, q_0, F)$における区別可能な対を再帰的に発見していくアルゴリズムである。

1. **基礎**：状態pが受理状態であり、状態qが受理状態でないならば対$\{p,q\}$は区別可能である。
2. **再帰**：状態pとqは、ある入力文字aに対し$r = \delta(p,a)$、$s = \delta(q,a)$であり、しかも$\{r,s\}$が区別可能であれば、p,qも区別可能である。

この判定が正しいのは以下の理由による。rとsが区別可能であることから、$\hat{\delta}(r,w)$と$\hat{\delta}(s,w)$のうちの一方のみが受理状態である文字列 w が存在する。そうすると、文字列awはpとqを区別する。$\{\hat{\delta}(p,aw), \hat{\delta}(q,aw)\}$は$\{\hat{\delta}(r,w), \hat{\delta}(s,w)\}$と同じ状態の対になるからである。

穴埋めアルゴリズムを用いることにより、二つの正規表現が同じ言語であるかどうかを判定する簡単な方法が得られる。言語 L と M が、二つの正規表現によって与えられているとする。まず、それぞれの表現を DFA に変換する[9]。その上で、L を受理する DFA の状態集合と M を受理する DFA の状態集合の集合和を状態集合として持つもう一つの DFA を想像する。定義上この DFA は二つの開始状態を持つことになるが、状態の同値性の判定が考察の対象となっている限りでは、開始状態の役割は不要である。任意の状態を開始状態として定めておいても差し支えない。

次は二つの元々の DFA で定義されていた開始状態同士について、それらが同値であるかどうかを、穴埋めアルゴリズムを用いて調べる。それらが同値であれば$L = M$であり、同値でないならば$L \neq M$である

4. 正規表現同一判定の効率化

4.1 正規表現の特徴量

正規表現の特徴量とは、正規表現が他の正規表現と区別する性質のことである。表 4.1 では本研究で使われる特徴量をまとめている。

要素数を$\#(\alpha)$で表し、文字列の最小長さを$|\alpha|_{min}$、最大値を$|\alpha|_{max}$で表す。要素数とは文字列の集合$L(\alpha)$の要素の数である。文字列の最小長さと最大長さは、文字列の集合の要素の中で最も短い要素の文字列の長さと最も長い文字列の長さである。

表 4.1　正規表現の特徴量

記号	意味
$\|\alpha\|_{min}$	$L(\alpha)$の文字列の最小長さ
$\|\alpha\|_{max}$	$L(\alpha)$の文字列の最大長さ
$\#(\alpha)$	言語$L(\alpha)$の要素数、$\#(\alpha) = \|L(\alpha)\|$

定理：同じ Σ 上の二つの正規表現αとβが同一であれば、下記の等式が成立する。

① $\#(\alpha) = \#(\beta)$

② $|\alpha|_{min} = |\beta|_{min}$

③ $|\alpha|_{max} = |\beta|_{max}$

④ 任意の$x \in L(\alpha)$であるとき、$x \in L(\beta)$である、逆に任意の$x \in L(\beta)$であるとき、$x \in L(\alpha)$である

4.2 特徴量の求め方

規表現の特徴量で同一判定をする前に特徴量を計算する必要がある。特徴量は構文木を用いて計算することができる。構文木とは構文解析の経過や結果を木構造で表したものである。木構造の各ノードは枝と葉に分かれ、葉のノードは値が必要である。枝のノードでは演算子とオペランドが必要である。オペランドは別の枝と葉のノードになる。正規表現の特徴量の計算に主に使われているオペレータは表 4.2 のようになる。

表 4.2　特徴量の計算に使われているオペレータ

正規表現	意味	構文木での表現
|	また	union
	連結	concat

[ ]	括弧内の 1 文字	anyof
{n} {n,m}	n 回繰り返し n 回以上 m 回以下	quant range, min, max
*	0 回以上繰り返し	star
+	1 回以上繰り返し	plus

4.2.1 言語の要素数の求め方

まずは構文木を用いて要素数を計算する方法から説明する。要素数を計算するとき、主に使われている記号の計算方法は下記の数式のようになる。

① anyof の計算

$\#([\ldots]) = |[\ldots]|$

② union の計算

$\#(\alpha|\beta) = \#(\alpha) + \#(\beta)$

③ concat の計算

$\#(\alpha\beta) = \#(\alpha) \times \#(\beta)$

④ quant の計算

$\#(\alpha\{m\}) = \#(\alpha)^m$
$\#(\alpha\{m,n\}) = \#(\alpha)^m + \#(\alpha)^{m+1} + \cdots + \#(\alpha)^n$

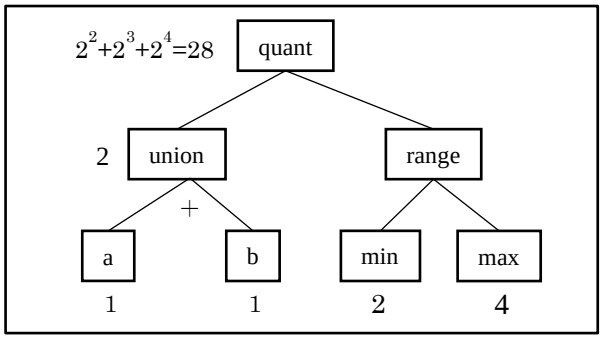

図 1 構文木における正規表現要素数の計算例

まず言語の要素数の例から説明する。

① 正規表現$\alpha = [a-c]\{2\}$であるとき

$L(\alpha) = \{aa, ab, ac, bb, ba, bc, ca, cb, cc\}$
$\#(\alpha) = |L(\alpha)| = 3^2 = 9$

② 正規表現$\alpha = [a-c][ab]$であるとき

$L(\alpha) = \{aa, ab, ba, bb, ca, cb\}$
$\#(\alpha) = |L(\alpha)| = 3 \times 2 = 6$

③ 正規表現$\alpha = [ab]\{3\}|[bc]\{2\}$であるとき

$L(\alpha) = \{aaa, aab, \cdots, bbb, bb, bc, cc, cb\}$
$\#(\alpha) = |L(\alpha)| = 2^3 + 2^2 = 12$

④ 正規表現$\alpha = a*$であるとき

$L(\alpha) = \{\varepsilon, a, aa, aaa \cdots\}$
$\#(\alpha) = |L(\alpha)| = \infty$

4.2.2 文字列の長さの求め方

次は構文木を用いて文字列の長さを計算する方法である。文字列の長さを計算するとき、主に使われている記号の計算方法は下記の数式のようになる。

① anyof の計算

$|[\ldots]|_{min} = |[\ldots]|_{max} = 1$

② union の計算

$|\alpha|\beta|_{min} = \min(|\alpha|_{min}, |\beta|_{min})$
$|\alpha|\beta|_{max} = \max(|\alpha|_{max}, |\beta|_{max})$

③ concat の計算

$|\alpha\beta|_{min} = |\alpha|_{min} + |\beta|_{min}$
$|\alpha\beta|_{max} = |\alpha|_{max} + |\beta|_{max}$

④ quant の計算

$|\alpha\{m\}| = n$
$|\alpha\{m,n\}|_{min} = n, |\alpha\{m,n\}|_{max} = m$

⑤ star の計算

$|\alpha*|_{min} = 0, |(\alpha)*|_{max} = \infty$

⑥ plus の計算

$|\alpha+|_{min} = |\alpha|_{min}, |\alpha+|_{max} = \infty$

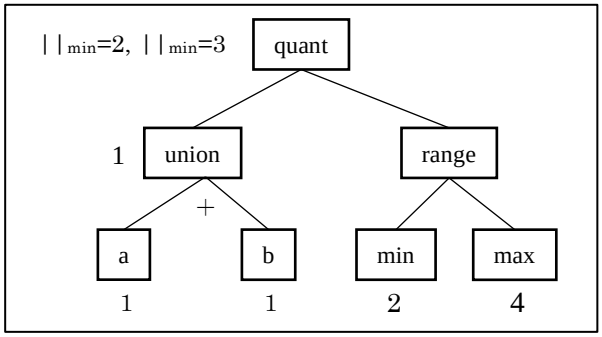

図 2 構文木における文字列長さの計算例

次は文字列の長さの例を説明する

① 正規表現$\alpha = [a-c]\{2\}$であるとき

$L(\alpha) = \{aa, ab, ac, bb, ba, bc, ca, cb, cc\}$
$|\alpha|_{min} = 2$、$|\beta|_{max} = 2$

② 正規表現$\alpha = [ab]\{2,3\}$であるとき

$L(\alpha) = \{aa, ab, ba, bb, aaa, aab, \cdots, baa, bab\}$

$|\alpha|_{min} = 2$、$|\alpha|_{max} = 3$

③ 正規表現$\alpha = [ab]\{3\}|[bc]\{2\}$であるとき

$L(\alpha) = \{aaa, aab, \cdots, bbb, bb, bc, cc, cb\}$
$|\alpha|_{min} = \min(|[ab]\{3\}|_{min}, |[bc]\{2\}|_{min}) = 2$
$|\alpha|_{max} = \max(|[ab]\{3\}|_{max}, |[bc]\{2\}|_{max}) = 3$

④ 正規表現$\alpha = a*$であるとき

$L(\alpha) = \{\varepsilon, a, aa, aaa \cdots\}$
$|\alpha|_{min} = 0$，$|\alpha|_{max} = \infty$

4.3 マッチングテストによる同一判定

マッチングテストによる同一判定である。正規表現αとβを与えられたとき、αにマッチする文字列（以降「αの文字列」とする）をテストデータして生成して、βにマッチするかをテストする。マッチしなければ同一ではないと判定する。すべてマッチした場合に、次はβの文字列を生成し、αにマッチするかをテストする。マッチしなければ、同一ではないと判定する。

入力：正規表現αとβ, テストに必要な文字列数 k
出力：αとβの同一判定の結果(True/False)
手順:
1. αの文字列をランダムに k 個生成
2. 生成された文字列がβにマッチするかをテストし、マッチしなければ、False を返して終了
3. βの文字列をランダムに k 個生成
4. 生成された文字列がαにマッチするかをテストし、マッチしなければ、False を返して終了
5. ここまで判断できなければ True を返して終了

リスト 1 マッチングテストによる同一判定

要素数 n の二つの正規表現 α と β の言語を L(α)と L(β)とする。L(α)と L(β)の要素の中で同じである要素を m 個とする。このとき、同一ではないが同一と判定する **False Positive な**確率 P が次のようになる。

$$P = \left(\frac{m}{n}\right)^{2k}$$

この式で示しているように、k が大きくなるほど確率 P は下がる。したがって、二つの違う正規表現が同一ではないという正確な判定をする確率は上がる。

しかし、この方法は全てのテストデータが正規表現にマッチしたとしても、二つの正規表現は同じであるとは判定できない。この方法は、二つの正規表現が違う場合しか判定できないが、判定する時間が既存の方法よりは短い。

5. 評価実験

本節では、提案の正規表現の同一判定手法を評価するための実験方法と実験結果について述べる。

5.1 実験方法

提案手法は非同一である場合に素早く判定でき、オートマトンによる同一判定を実行しなくて済むが、最終判定ができないケースもある。提案手法と既存のオートマトンによる同一判定と組み合わせで実験を行う必要がある。まず、オートマトンによる同一判定のみ実行し、時間を測る。次はオートマトンによる同一判定を実行する前に提案の方法を入れて、時間を測り、時間が節約できて効率化が実現できるかどうかを評価する。

実験に使う正規表現は四つのセットを用意する。(A) 正規表現セット、(B) 同一正規表現セット、(C) 非同一正規表現セット、(D) 部分同一正規表現セット（4 割が非同一とする）である。各セットの中には事前に用意した正規表現が入っている。この四つのセットを使って同一判定を行う。

作った正規表現の要素数は多くないので、今回はマッチングテストによる同一判定で生成する文字列数 k=5 にして実験を行う。正規表現セット A と他の三つのセットから一セットずつ同一判定をして時間を測る。同一判定はオートマトンによる同一判定のみ実行し、提案方法とオートマトンによる同一判定を一つずつ合わせて実行して、最後は四つの方法とオートマトンによる同一判定を合わせて実行する。実験をするために使う変数について説明する。

表 5.1 パラメータの説明

パラメータ	意味	値
loop	設定した同一判定の繰り返し回数	100
step	繰り返し回数の増加量	100
max_loop	最大繰り返し回数	1000
config	最小長さ、最大長さ、要素数、マッチングテストの実行制御	1：実行 0：実行しない

ここで loop は設定した同一判定の繰り返し回数である。例えば config が(0,0,0,1)の場合はマッチングテストによる同一判定をした後オートマトンによる同一判定をするが、このセットを繰り返す意味である。最初は設定した同一判定の回数は 100 回から始め、次から回数を 100 回ずつ増えて同一判定を行う。例えば、100 回、200 回、300 回である。このように繰り返し回数が最大繰り返し回数と同じであれば処理を停止する。ここで、config の括弧の中の四つの値は、それぞれ、文字列の最小長さによる同一判定、文字列の最大長さによる同一判定、言語の要素数による同一判定、

マッチングテストによる同一判定を実行するかどうかを表している。

5.2 実験結果

実験用プログラムは Perl で開発している。正規表現の構文解析に Regexp::Parser[16]が使われた。マッチングテストでは String::Random[17]というパッケージを使って正規表現からランダムに文字列を生成できる。実験結果は図 3～図 5 に示している。

まず、同一正規表現セット B と元正規表現セット A の同一判定の実験を行った。図 3 では同一正規表現セットの場合の結果を比較した結果を示している。

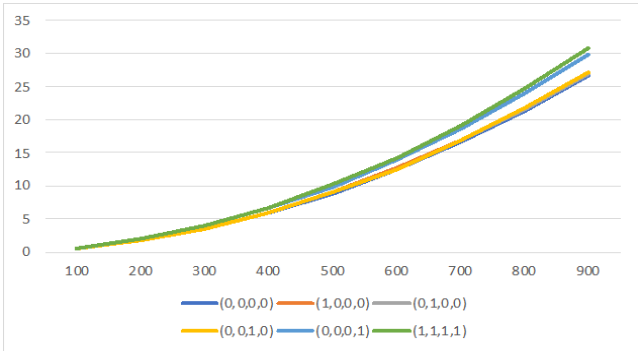

図 3 同一正規表現セットの評価結果

提案方法は同一である二つの正規表現は判定できないため、オートマトンによる同一判定のみ実行する場合とほぼ同じである。

つぎに、非同一正規表現セット C と元正規表現セット A の同一判定の実験を行った。図 4 は非同一正規表現セットの場合の結果を比較した結果を示している。

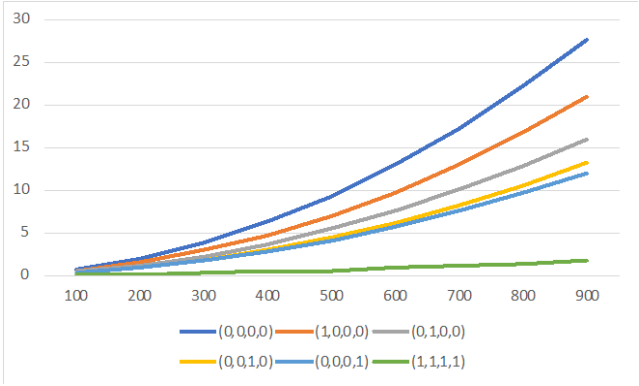

図 4 非同一正規表現セットの評価結果

非同一正規表現セットを使う場合は、非同一と判定されたら、オートマトンによる同一判定が実行せずに済むため、実行時間が明らかに短くなっている。特に 4 つの方法を合わせて実行する場合は最も短くなっており、提案方法が有効であることを示している

最後に、部分同一正規表現セットD と元正規表現セット A の同一判定の実験を行った。図 5 は部分同一正規表現セットの場合の結果を比較した結果を示している。

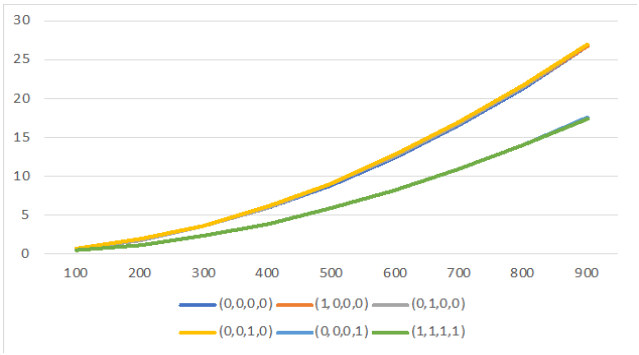

図 5 部分同一正規表現セットの評価結果

部分同一正規表現に対しても、提案手法をすべて利用する場合は効率化が確認できた。特にマッチングテストが一番寄与するところが大きいことも確認できた。

6. おわりに

正規表現は書き方が違っても実質同じであることがありうる。正規表現の同一判定のため、既存手法ではオートマトンを利用することが一般的であるが、状態数やΣの大きさに依存しており計算コストが高い問題がある。本研究では、正規表現の特徴量による同一判定、マッチングテストによる同一判定を提案し、同一判定の効率化を図った。

正規表現の特徴量がオートマトンの状態数や記号列のサイズに依存せず、非同一判定の効率化に寄与できる。提案手法を評価するために、任意の正規表現に対応する言語の最小長さ、最大長さ、要素数を求めるアルゴリズムを開発し、特徴量、マッチングテストによる同一判定を Perl で実装し、評価実験を行った。その結果、非同一の割合が高ければ、有限オートマトンを使わず、判定でき、効率化を確認できた。

今後の課題として、マッチングテストにおいて、特徴量に基づき、必要な文字列数 k の最適化や、実験データの増加などが挙げられる。

謝辞 本研究は令和 2 年度 KSU 基盤研究費による支援を頂いており、ここで謹んで感謝の意を表する.

参考文献

[1]. J.E. Hopcroft, R. Motwani, J. D. Ullman, Introduction to Automata Theory, Languages and Computation. Addison Wesley (2000).
[2]. John E. Hopcroft. An n log n algorithm for minimizing states in a finite automaton. Technical Report. Stanford University, Stanford, CA, USA. (1971)
[3]. J. E. Hopcroft, R. M. Karp, A linear algorithm for testing equivalence of finite automata, Technical Report 71-114, University of California (1971).
[4]. V. M. Antimirov, P. D. Mosses, Rewriting extended regular expressions. In: G. Rozenberg and A. Salomaa, editors:

Developments in Language Theory, World Scientific, pp.195-209(1994).
[5]. V. M. Antimirov, Partial derivatives of regular expressions and finite automaton constructions.Theoret. Comput. Sci.155, pp.291-319 (1996).
[6]. M. Almeida, N. Moreira, R. Reis, Antimirov and Mosses's rewrite system revisited. In: O. Ibarraa & B. Ravikumar, editors: CIAA 2008, LNCS 5448, Springer-Verlag, pp.46-56 (2008).
[7]. M. Almeida, N.Moreira , R. Reis , Enumeration and generation with a string automata representation, Theoret. Comput. Sci. 387, pp. 93-102 (2007).
[8]. M. Almeida, N. Moreira, R. Reis, Testing the Equivalence of Regular Languages, Journal of Automata, Languages and Combinatorics 15(2010)1/2, pp.7-25.
[9]. J.ホップクロフト, R.モトワニ, J.ウルマン, オートマトン言語理論 計算論 I [第２版], サイエンス社, 2003.
[10]. 杉山 貴章, 反復学習ソフト付き正規表現書き方ドリル, 技術評論社(2011).
[11]. 宮前 竜也, [改訂新版]正規表現ポケットリファレンス, 技術評論社, 2015.
[12]. 新屋 良磨, 鈴木 勇介, 高田 謙, 正規表現技術入門, 技術評論社, 2015.
[13]. 佐藤 竜一, 正規表現辞典 改訂新版, 翔泳社, 2018
[14]. 小倉 久和, 形式言語と有限オートマトン入門, コロナ社, 1996.
[15]. Jeffrey E. F. Friedl, 詳説 正規表現 第 3 版, オライリー・ジャパン, 2008.
[16]. Todd Rinald, Regexp::Parser - base class for parsing regexes (2020.12), https://metacpan.org/pod/Regexp::Parser
[17]. String::Random-Perl module to generate random strings based on a pattern, (2020.12), https://metacpan.org/pod/String::Random
[18]. Regexp::ERE-extended regular expressions and finite automata, (2020.12), https://metacpan.org/pod/Regexp::ERE