UNIVERSITÄT TRIER

Top-Down Construction of Finite Automata from Regular Expressions

Stefan Gulan
FB IV, Informatik
gulan@uni-trier.de

Henning Fernau
FB IV, Informatik
fernau@uni-trier.de

TRIERER FORSCHUNGSBERICHTE
INFORMATIK / MATHEMATIK

NO. 08-7

Juli 2008

Abstract

We consider the construction of finite automata from their corresponding regular expressions by a series of digraph-transformations along the expression's structure. Each intermediate graph represents an extended finite automaton accepting the same language. The character of our construction allows a fine-grained analysis of the emerging automaton's size, eventually leading to an optimality result, i.e., a tight bound.

1 Introduction

Regular expressions provide a description of regular languages in a manner convenient for the human reader. On the machine level, however, the most appropriate representation is arguably that of finite automata. Thus, considerable effort has been put into ways of constructing automata describing the same language as a given expression. All algorithms known to the authors work by either incorporating the expression's syntactic structure into the state graph of the emerging automaton [OF61, Kle65, Tho68, SSS88, IY03] or by looking for first-time occurrences of symbols in subexpressions [Glu61, MY60, BS86]. The first kind of construction generally results in an NFA with ϵ-transitions (ϵNFA, for short), the latter produces no such transitions and may even provide a DFA. An exhaustive overview, structured by a more subtle, categorization is given in [Wat94].

Our construction yields an ϵNFA. No tight bound for the size of such an automaton representing a given expression has been published yet. Ilie & Yu [IY03] came pretty close, proving a lower bound of $\frac{4}{3}$ times the size of a given expression while constructing an ϵNFA smaller than $\frac{3}{2}$ times the expression length. We close this gap by raising the lower bound and giving a construction reaching that bound in the worst case. Unfortunately, there are plenty of definitions of the sizes of finite automata and regular expressions. This holds especially for regular expressions, however, in [EKSW05], the authors show how the different values relate to each other. For comparability, we stick by the definition given in [IY03].

The algorithm presented in this paper is basically an extension to the one in [OF61], which is, together with a variation of Thompson's algorithm in [Wat94], the only top-down algorithm among a variety of bottom-up procedures. It turns out that the top-down character is very helpful in the analysis, since it allows systematic construction of an expression yielding the worst ratio of automaton-to-expression sizes. This construction relies on extremal combinatorial arguments for inferring structural properties of a worst-case input. To our knowledge this is a novel approach to this kind of problem.

2 Preliminaries

Enclosing braces for singleton sets will be omitted. Let $\mathcal{A}$ be a finite set of symbols, called *alphabet*, the elements of $\mathcal{A}\cup\epsilon$ will be called *literals*. The set of regular expressions over $\mathcal{A}$, denoted $Reg(\mathcal{A})$, is the closure of $\mathcal{A}\cup\epsilon$ under product $\bullet$, sum $+$ and Kleene-star *. Operator precedence is $*, \bullet, +$. We will casually speak of *expressions* only. In the following, α and β will always be expressions. The regular language expressed by α is denoted $L(\alpha)$. We will call α and β *equivalent*, denoted $\alpha\equiv\beta$, if $L(\alpha)=L(\beta)$. The number of products (sums, stars) in α will be denoted $|\alpha|_\bullet$ ($|\alpha|_+$, $|\alpha|_*$). Likewise, the number of literals in α, counted with multiplicity, will be denoted $|\alpha|_\mathcal{A}$. The *operator complexity* of α is defined as $|\alpha|_{op}:=|\alpha|_\bullet+|\alpha|_++|\alpha|_*$. The *size* of an expression is defined as $|\alpha|{:=}|\alpha|_{\mathcal{A}}{+}|\alpha|_{op}$. We call α *complex*, if $|\alpha| \geq 2$. The set of subexpressions of α will be denoted $sub(\alpha)$.

Both iterated products and sums will be denoted as is common in arithmetic, defining

$$\prod_{i=1}^{n} \alpha_i := \alpha_1 \bullet \alpha_2 \bullet \ldots \bullet \alpha_n \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i := \alpha_1 + \alpha_2 + \ldots + \alpha_n$$

Each α_i as above will be called an *operand* to the product or sum. An iterated product (sum) which is not operand to a product (sum) itself, will be called *maximal*. If all operands in a maximal product (sum) are starred, it will be called *star-maximal*.

An *extended finite automaton*, short *EFA*, is a 5-tuple $E{=}(Q, \mathcal{A}, \delta, q_0, F)$, where $q_0{\in}Q$, $F{\subseteq}Q$, and $\delta{\subset}Q{\times}Reg(\mathcal{A}){\times}Q$. This renders conventional FAs a special case of EFAs. An EFA is called *normalized*, if $|F|{=}1$. A pair $(q, w) \in Q{\times}\mathcal{A}^*$ is called *configuration* of E, valid changes in E's configuration are denoted by $\vdash$, writing $(q, vw) \vdash (q', w)$ if $(q, \alpha, q'){\in}\delta$ and $v{\in}L(\alpha)$. The language accepted by an EFA E is $L(E){=}\{w|(q_0, w) \vdash^* (q_f, \epsilon), q_f \in F\}$, where $\vdash^*$ is the reflexive-transitive closure of $\vdash$.

The class of regular languages is not extended by allowing regular expressions as labels in automata, see [Woo87] for a proper introduction. The size of an EFA E is $|E|{:=}|Q|{+}|\delta|$. The sets of transitions leaving and reaching some $q{\in}Q$ are given by $q^+{:=}\delta{\cap}(q{\times}Reg(\mathcal{A}){\times}Q)$ and $q^-{:=}\delta{\cap}(Q{\times}Reg(\mathcal{A}){\times}q)$, respectively. The notion is extended to sets $S{\subseteq}Q$, thus $S^+ = \bigcup_{s\in S} s^+$, resp. S^-. We call $\gamma{\subseteq}\delta$ an ϵ-cycle in E, if $\gamma = \{(q_i, \epsilon, q_{i+1})|1{\leq}i{\leq}|\gamma|{-}1\} \cup \{(q_{|\gamma|}, \epsilon, q_1)\}$.

Let A be a FA generated from α by some algorithm $\mathcal{C}$. We call $\frac{|A|}{|\alpha|}$ the *conversion-ratio* of $\mathcal{C}$ with respect to α. The maximal conversion-ratio of $\mathcal{C}$ with respect to any expression, will simply be called conversion-ratio of $\mathcal{C}$. An expression reaching this bound is said to be *worst-case*.

3 A Lower Bound

First we improve on a lower bound for *any* construction of FAs from expressions, given by Ilie & Yu in [IY03], by a slight variation of their argument. To this end, a property of digraphs is shown, in which we refer to both vertices and arcs as *elements*.

Proposition 3.1. *Consider a digraph (V, A). Let L, R be nonempty, disjoint subsets of V such that*

1. *there is a path from each $l \in L$ to each $r \in R$,*
2. *there is no path connecting any two vertices $l, l' \in L$ or any $r, r' \in R$.*

Then at least $\min\{|L||R|, |L|{+}|R|{+}1\}$ *elements are necessary to realize these paths.*

Proof. Two cases need to be considered:

1. There is no vertex on any path connecting l with r. This can only be realized with $|L||R|$ arcs, by pairwise connections.
2. There is at least one vertex b on a path connecting $l_b{\in}L$ with $r_b{\in}R$, this path contains at least 3 elements. To connect l_b with the vertices of $R\backslash r_b$ at least $|R|{-}1$ further arcs are necessary. An additional $|L|{-}1$ arcs are leaving the vertices of $L\backslash l_b$. These numbers total to $|L|{+}|R|{+}1$. □

Next we show the actual lower bound. Both states and transitions of an FA A will be called elements, the number of elements is simply $|A|$.

Theorem 3.1. *Let $x_{i,j}$ be distinct literals, consider the expression*

$$\begin{aligned}\alpha &= \prod_{i=1}^{n}(x_{2i-1,1}^{*}+x_{2i-1,2}^{*})(x_{2i,1}^{*}+x_{2i,2}^{*}+x_{2i,3}^{*})\\ &= (x_{1,1}^{*}+x_{1,2}^{*})(x_{2,1}^{*}+x_{2,2}^{*}+x_{2,3}^{*})\ldots(x_{2n-1,1}^{*}+x_{2n-1,2}^{*})(x_{2n,1}^{*}+x_{2n,2}^{*}+x_{2n,3}^{*})\end{aligned}$$

Any normalized automaton A satisfying $L(A)=L(\alpha)$ has at least size $22n+1$.

Proof. A has to provide disjoint cycles accepting some $x_{i,j}^{k}$, otherwise the relative order of the $x_{i,j}$ could be violated. This calls for at least a state $q_{i,j}$ and a transition per $x_{i,j}$, or $10n$ elements. Further, there must be paths from $q_{i,j}$ to $q_{i+s,k}$ for reasonable $i,j,k,s \geq 1$. This requires at least the existence of paths from $q_{i,j}$ to $q_{i+1,k}$. The order of literals in α disallows paths from $q_{i,j}$ to $q_{i,k}$ where $k\neq j$. By Prop. 3.1, each two layers $q_{i,j}, q_{i+1,k}$ are connected by at least 6 elements, thus $2n-1$ layers require at least $12n-6$ more elements. Finally, 7 elements result from requiring A to be normalized. The total number of elements, i.e., the size of A is thus at least $22n+1$. $\square$

For the following, note that α from Thm. 3.1 has size $15n-1$.

Corollary 3.1. *The conversion-ratio of any algorithm converting expressions to normalized FAs is bounded from below by*

$$\frac{|A|}{|\alpha|} \geq \frac{22n+1}{15n-1} > \frac{22}{15}+\frac{1}{|\alpha|} = 1.4\bar{6}+\frac{1}{|\alpha|}$$

4 Construction

The idea is to expand an initial EFA according to the structure of the expression, by introducing as few states and transitions as possible, while decomposing transition labels. Certain substructures in the expanded automata will be replaced by smaller equivalents. This is done until an ϵNFA emerges, i.e., there are no more complex labels.

Definition 4.1 (Expansion)**.** *Let $E=(Q,\mathcal{A},\delta,q_0,F)$ be an EFA with a complex labeled transition t. We call an EFA $E'=(Q',\mathcal{A},\delta',q_0,F)$ the* expansion *of E, if it is derived from E according to t's label as follows:*

- *if* $t=(p,\alpha\beta,q)$ *then* $Q'=Q\dot{\cup}p'$, $\delta'=\delta\setminus t\cup\{(p,\alpha,p'),(p',\beta,q)\}$
- *if* $t=(p,\alpha+\beta,q)$ *then* $Q'=Q$, $\delta'=\delta\setminus t\cup\{(p,\alpha,q),(p,\beta,q)\}$
- *if* $t=(p,\alpha^{*},q)$, *we distinguish several cases*

  $*0$: *if* $p=q$, *replace* α^{*} *with* α,
  let $Q'=Q$, $\delta'=\delta\setminus t\cup(q,\alpha,q)$

  $*1$: *if* $|p^{+}|=|q^{-}|=1$, *merge* q *into* p:
  let $Q'=Q\setminus q$, $\delta'=\delta\setminus(q^{+}\cup q^{-})\cup\{(p,\gamma,r)|(q,\gamma,r)\in\delta\}\cup(p,\alpha,p)$

  $*2$: *if* $|p^{+}|>1,\ |q^{-}|=1$, *introduce a loop in* q:
  let $Q'=Q$, $\delta'=\delta\setminus t\cup\{(p,\epsilon,q),(q,\alpha,q)\}$

  $*3$: *if* $|p^{+}|=1,\ |q^{-}|>1$, *introduce a loop in* p:
  let $Q'=Q$, $\delta'=\delta\setminus t\cup\{(p,\alpha,p),(p,\epsilon,q)\}$

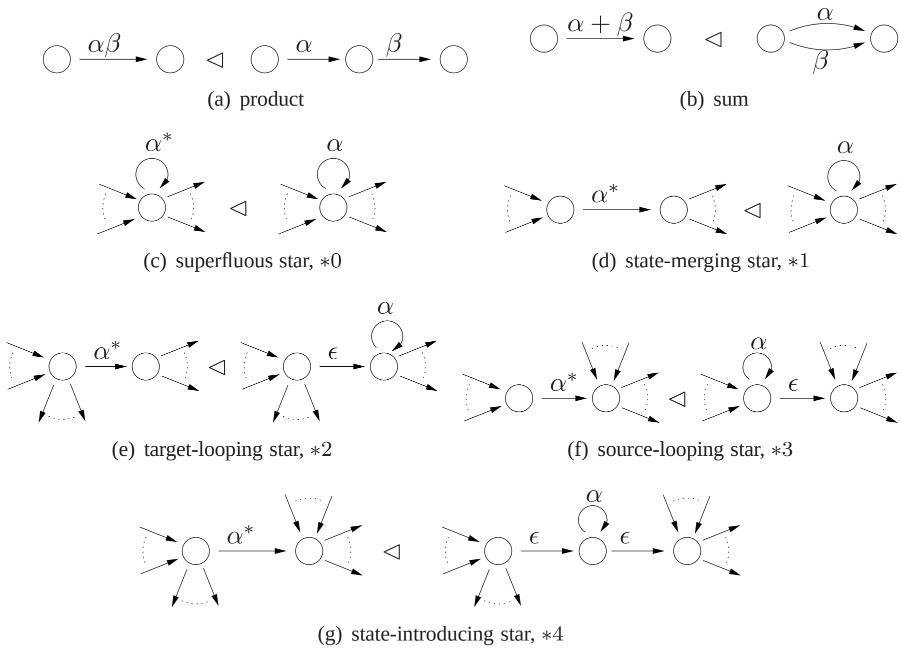


Figure 1: Expansions of complex labeled transitions.

$*4$*: if* $|p^+| > 1,\ |q^-| > 1$*, introduce a new state* p'*:*
let $Q' = Q\dot{\cup}p'$*,* $\delta' = \delta \setminus t \cup \{(p, \epsilon, p'), (p', \alpha, p'), (p', \epsilon, q)\}$

Cases are sketched in Fig. 1. Expansions will be denoted relational, writing $E \lhd_t E'$ if E' results from expansion of t in E. Occasionally we will also write $\lhd_\bullet, \lhd_+, \lhd_*$ or simply $\lhd$, if t or its root operator are irrelevant. This might be formalized as $\lhd = \lhd_\bullet \cup \lhd_+ \cup \lhd_*$. The n-fold iteration of $\lhd$ will be denoted $\lhd^n$, thus if $E \lhd^n E'$ there is a series of EFAs E_i, $0 \leq i \leq n$, such that $E = E_0$, $E_i \lhd E_{i+1}$, $E_n = E'$. Usually we refer to $\lhd_{(q,\alpha,q')}$ by mentioning α's operator, e.g, '$\bullet$-expansion'. Distinct $*$-expansions will be referred to as '$*0$-expansion' to '$*4$-expansion' according to Def. 4.1.

Definition 4.2. *Let $\mathcal{A}$ be the least alphabet satisfying $\alpha \in Reg(\mathcal{A})$. The EFA $A^0_\alpha = (\{q_0, q_f\}, \mathcal{A}, (q_0, \alpha, q_f), q_0, q_f\})$ is called the* primal EFA representing α*. We denote by A^i_α any automaton satisfying $A^0_\alpha \lhd^i A^i_\alpha$.*

Thus, A^i_α denotes any EFA derived from the primal automaton representing α in a series of i expansions. Note that generally, A^i_α is not unique. However, a most useful property of $\lhd$ is that the order of expansion is irrelevant, or formally:

Lemma 4.1. *$\lhd$ is locally confluent modulo isomorphism, i.e., if $A \lhd B_1$ and $A \lhd B_2$, then $\exists C_1, C_2 : B_1 \lhd C_1$ and $B_2 \lhd C_2$ and $C_1 \equiv C_2$.*

Proof. Let $A \lhd_{t_1} B_1$ and $A \lhd_{t_2} B_2$. First, assume one of the transitions is labeled by either a product or a sum:

- Let $t_1 = (q, \alpha \bullet \beta, q')$. Upon expansion a bridge-state q'' will be introduced, however the number of arcs leaving and reaching q and q' will remain constant. The structure of A will change insofar as that an arc will be elongated. Since any $\lhd_{t_2}$ will at most have the effect on t_1 that one of its states might be renamed (upon $*1$-expansion), the order of $\lhd_{t_1}, \lhd_{t_2}$ is irrelevant.

- If $t_1{=}(q, \alpha{+}\beta, q')$, informal reasoning is that an arc is merely doubled. Looking at Def. 4.1, the booleans $q^+{>}1$ etc. are not changed by such an operation.

Now let both t_i be star-labeled. Note that the statement is trivial, if expansions take place in 'different parts' of the EFA, so let t_1, t_2 share at least a common state. If the transitions are parallel, both will be $*4$-expanded anyway. Further, $*0$-expansion does not change the structure of the state-graph at all, i.e., neither of t_1, t_2 is a loop. So assume $t_1 = (p, \alpha^*, q)$, $t_2 = (q, \alpha^*, r)$ where $p \neq q \neq r$. Some of the possible combinations are shown in Fig. 4, the remaining are a simple exercise. $\square$

Corollary 4.1. $\lhd$ *is confluent.*

Proof. Since $\lhd$ is terminating, the claim follows from Lem. 4.1. Detailed proof of this argument can be found, e.g., in [Hue80]. $\square$

We introduce two further conversions of different nature, altering EFAs with respect to ϵ-labeled substructures.

Definition 4.3 (State-Elimination)**.** *Let $E{=}(Q, \mathcal{A}, \delta, q_0, F)$ be an EFA, $q \in Q \setminus F$ such that*

- $|q^+||q^-| \leq |q^+|{+}|q^-|$
- *all labels in q^+ or in q^- (or both) are ϵ*

*Let $\delta'{=}\delta\backslash(q^+{\cup}q^-) \cup \{(q', ab, q'')|(q', a, q){\in}q^-,\ (q, b, q''){\in}q^+\}$, then the q-*reduct *of E is defined as $E' = (Q\backslash q, \mathcal{A}, \delta', q_0, F)$, and we write $E {\rhd_q} E'$.*

State-elimination can be applied to q, if all leaving or all reaching transitions (or both) are labeled ϵ and if either of $|q^+|, |q^-|$ is 1, or both are 2. Though state-eliminations reduces the size of an EFA by either 1 or 2 its effects will turn out to be crucial in the proof of Thm. 5.2.

Definition 4.4 (ϵ-Cycle-Elimination)**.** *Let $E{=}(Q, \mathcal{A}, \delta, q_0, F)$ be an EFA, γ an ϵ-cycle connecting the set of states Q_γ, then the EFA $E'{=}(Q', \mathcal{A}, \delta', q_0, F)$ is called the γ-*reduct*, written $E {\rhd_\gamma} E'$, if*

- $Q' = Q \setminus Q_\gamma \cup q_\gamma$ *where* $q_\gamma \notin Q$
- $\delta' = \delta\backslash(\gamma{\cup}Q_\gamma^+{\cup}Q_\gamma^-){\cup}\{(q_\gamma, \alpha, q')|(q, \alpha, q'){\in}Q_\gamma^+\}{\cup}\{(q', \alpha, q_\gamma)|(q', \alpha, q){\in}Q_\gamma^-\}$

Note that both eliminations strictly reduce the size of an EFA without re-introducing complex labels. Exhaustive application of these transformations to a primal automaton yields an ϵNFA.

Proposition 4.1 (Acceptance-Invariance)**.** *Let E, E' be EFAs. Then $L(E) = L(E')$, if $E \lhd E'$ or $E \rhd_q E'$ or $E \rhd_\gamma E'$.*

A schematic algorithm showing how to convert from an expression to a corresponding ϵNFA is given in Alg. 1. The algorithm is intentionally crude and has unfortunate time-complexity — however, runtime-analysis and implementation details lie not within the scope of this report.

Algorithm 1 RegEx $\rightarrow$ ϵNFA

$A \leftarrow A^0_\alpha$
while A is not an NFA **do**
  choose a complex-labeled transition t in A
  let $A \lhd_t A'$
  if $\lhd_t$ introduced some $e = (q, \epsilon, q')$ **then**
    if q can be eliminated **then**
      let $A' \rhd_q A''$
      $A' \leftarrow A''$
    if q' can be eliminated **then**
      let $A' \rhd_{q'} A''$
      $A' \leftarrow A''$
    if e is part of some ϵ-cycle γ **then**
      let $A' \rhd_\gamma A''$
      $A' \leftarrow A''$
  $A \leftarrow A'$
end while

	$\lhd_\bullet$	$\lhd_+$	$\lhd_{*0}$	$\lhd_{*1}$	$\lhd_{*2}$, $\lhd_{*3}$	$\lhd_{*4}$	$\rhd_\gamma$	$\rhd_q$
$\Delta(\lvert Q\rvert)$	1	0	0	-1	0	1	-($\lvert\gamma\rvert - 1$)	-1
$\Delta(\lvert\delta\rvert)$	1	1	0	0	1	2	-$\lvert\gamma\rvert$	-1 or 0

Table 1: Number of elements introduced (i.e., removed, if negative) upon expansion and elimination, broken down to states and transitions.

Clearly, there is no infinite series of expansions and reductions of an EFA, both operations either reduce the overall length of labels or the number of states and transitions. Thus, Alg. 1 is terminating. By Prop. 4.1 it is also partially correct, thus the algorithm is totally correct.

5 Analysis

Let A_α denote an ϵNFA constructed by our algorithm from A^0_α. We start by bounding $|A_\alpha|$ from above. To this end, we refine the definition of $|\alpha|_*$. Let $|\alpha|_{*i}$ denote the number of stars in α which will be $*i$-expanded. Clearly, $|\alpha|_* = \sum_{0 \le i \le 4} |\alpha|_{*i}$.

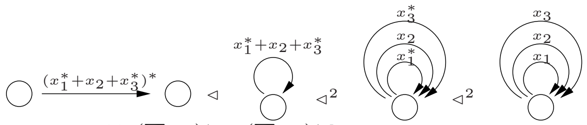
Theorem 5.1. *The size of an automaton built from α by our algorithm is bounded by*

$$|A_\alpha| \le |\alpha| + 2|\alpha|_{*4} - |\alpha|_+ + 2$$

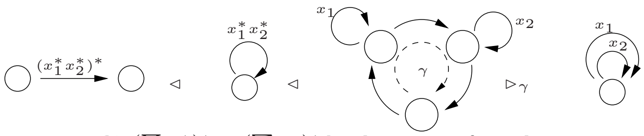
The bound is tight iff neither elimination nor $*0, *1$-expansion is applied.

Proof. A^0_α has size 3. The number of elements introduced upon expansion is determined by $|\alpha|_\bullet, |\alpha|_+, \ldots$, weighted by the entries in Tab. 1. Using $|\alpha|_{\mathcal{A}} = |\alpha|_\bullet + |\alpha|_+ + 1$ and $|\alpha| = |\alpha|_\bullet + |\alpha|_+ + |\alpha|_{*0} + \ldots + |\alpha|_{*4} + |\alpha|_{\mathcal{A}}$, this yields:

$$\begin{aligned} |A_\alpha| &\le 2|\alpha|_\bullet + |\alpha|_+ - |\alpha|_{*1} + |\alpha|_{*2,3} + 3|\alpha|_{*4} + 3 \\ &= |\alpha| + |\alpha|_\bullet - |\alpha|_{*0} - 2|\alpha|_{*1} + 2|\alpha|_{*4} - |\alpha|_{\mathcal{A}} + 3 \\ &\le |\alpha| + |\alpha|_\bullet + 2|\alpha|_{*4} - |\alpha|_{\mathcal{A}} + 3 \\ &= |\alpha| + 2|\alpha|_{*4} - |\alpha|_+ + 2 \end{aligned}$$



(a) $(\sum \alpha_i)^* \equiv (\sum \overline{\alpha_i})^*$ by $*0$-expansions.



(b) $(\prod \alpha_i^*)^* \equiv (\sum \alpha_i)^*$ by elimination of ϵ-cycles.

Figure 2: Transformations respect the equivalences given in Prop. 5.1 (ϵ-labels are omitted).

The first inequality results from state- and ϵ-cycle eliminations, the second from $*0$- and $*1$-expansions, thus equality holds in absence of these transformations. □

Proposition 5.1. *Both sides in each of the following equivalences will be expanded to the same (sub)automaton:*

$$(\alpha^*)^* \equiv \alpha^* \quad \textit{and} \quad (\sum \alpha_i)^* \equiv (\sum \overline{\alpha_i})^* \quad \textit{and} \quad (\prod \alpha_i^*)^* \equiv (\sum \alpha_i)^*$$

where $\overline{\alpha_i} = \beta_i$*, if* $\alpha_i = \beta_i^*$ *and* α_i *otherwise.*

Proof. The first two equivalences are realized by $*0$-expansion, the third by ϵ-cycle-elimination. Examples are given in Fig. 2. □

Corollary 5.1. *Let* α *be worst-case, then*

1. *The conversion-ratio for* α *is*

$$\frac{|A_\alpha|}{|\alpha|} = 1 + \frac{2|\alpha|_{*4} - |\alpha|_+ + 2}{|\alpha|}$$

2. $|\alpha|_{*0} = |\alpha|_{*1} = 0$*, further both a sum with starred operands and a maximally starred product are not starred themselves.*

Proof. The first item is a trivial consequence of Thm. 5.1. The second follows from Prop. 5.1 and Thm. 5.1. □

We proceed with a series of results, each putting additional constraints to the structure of a worst-case expression. Almost all proofs work by a line of argumentation that is common in extremal combinatorics: assume α is worst-case (clearly, such an expression must exist), i.e., extremal with respect to conversion-ratio, then infer some further property by contradicting extremality of α.

Proposition 5.2. *A worst-case expression contains stars.*

Proof. Let α be worst-case with $|\alpha|_*=0$. Cor. 5.1 implies $\frac{|A_\alpha|}{|\alpha|} \leq 1+\frac{2}{|\alpha|}$, the right-hand side of which drops below 1.4, if $|\alpha| \geq 5$. Since by Cor. 3.1, the conversion-ratio is bounded from below by $1.4\bar{6}$, the assumption $|\alpha|_*=0$ is wrong, if α is worst-case. □

Lemma 5.1. *Let γ^* be a proper subexpression of α. Then γ^* will be $*4$-expanded iff*

- *it is operand to a sum, which is not starred itself, or*
- *wlog. it occurs rightmost in a star-maximal product.*

Proof. The first case is clear by looking at the expansion of some $\gamma^*{+}\beta$. If γ^* is an infix, say, $\alpha_1\gamma^*\alpha_2$, we distinguish 3 cases: If both α_i are non-starred, γ^* will be $*1$-expanded. If only one of the α_i is non-starred, then γ^* can be $*2$- or $*3$-expanded by introducing a loop at the state incident to the transition labeled with the non-starred α_i. Finally, if both α_i are starred, we can by confluence assume that expansions will be applied from left to right. Then, every starred factor will be $*2$-expanded until the final one necessitates $*4$-expansion. This embraces all possible cases, giving both directions of the statement. $\square$

Lemma 5.2. *Let α be worst-case, assume $\gamma^*{\in}sub(\alpha)$ is $*4$-expanded. Then γ^* is operand to a sum.*

Proof. By Lem. 5.1, γ^* is either operand to a sum or rightmost in a star-maximal product. Assume the latter, thus $\pi = \pi_1^* \bullet \ldots \bullet \pi_{n-1}^* \bullet \gamma^*$. Construct α' from α by replacing π with $\sigma = \pi_1^*{+} \ldots + \pi_{n-1}^*{+}\gamma^*$. Then $|\alpha|{=}|\alpha'|$, however $2|\alpha'|_{*4}{-}|\alpha'|_+ = 2|\alpha|_{*4}{-}|\alpha|_+{+}n{-}1$. Since by Prop. 5.1 π is not starred in α, the stars in σ will not accidentally become $*0$. By Cor. 5.1, $\frac{|A_{\alpha'}|}{|\alpha'|} > \frac{|A_\alpha|}{|\alpha|}$, thus α is not worst-case. Therefore γ^* is necessarily operand to a sum. $\square$

The interrelation between sums and stars in a worst-case expression is tightened in

Lemma 5.3. *Let α be worst-case. Then*

1. *every starred subexpression in α is operand to a sum and*
2. *all operands in a maximal sum are starred.*

Proof.

1. Assume $\gamma^*{\in}sub(\alpha)$ will not be $*4$-expanded. Construct α' from α by replacing γ^* with γ. Since $|\alpha'|{=}|\alpha|{-}1$, yet $|\alpha'|_{*4}{=}|\alpha|_{*4}$, Cor. 5.1 again yields $\frac{|A_{\alpha'}|}{|\alpha'|} > \frac{|A_\alpha|}{|\alpha|}$, thus α is not worst-case. Therefore each star in a worst-case expression is subject to $*4$-expansion, thus by Lem. 5.2 operand to a sum.

2. Let $\sum \sigma_i$ be maximal with some σ_j unstarred, i.e., a product. Construct α' from α by replacing σ_j with σ_j^*. Then $|\alpha'| = |\alpha| + 1$, $|\alpha'|_{*4} = |\alpha|_{*4} + 1$ and by Cor. 5.1, $|A_{\alpha'}| = |A_\alpha| + 2$. Now

$$\frac{|A_{\alpha'}|}{|\alpha'|} = \frac{|A_\alpha| + 2}{|\alpha| + 1} > \frac{|A_\alpha|}{|\alpha|} \quad \text{iff} \quad |A_\alpha| < 2|\alpha|$$

   We proceed similar to the proof of Thm. 5.1, additionally using that the previous item implies $|\alpha|_{*4} \leq 2|\alpha|_+$:

$$\begin{aligned}
|A_\alpha| &\leq 2|\alpha|_\bullet + |\alpha|_+ - |\alpha|_{*1} + |\alpha|_{*2,3} + 3|\alpha|_{*4} + 3 \\
&= 2|\alpha| - |\alpha|_+ - 3|\alpha|_{*1} - |\alpha|_{*2,3} + |\alpha|_{*4} + 3 - 2|\alpha|_{\mathcal{A}} \\
&= 2|\alpha| - 2|\alpha|_+ - |\alpha|_\bullet - 3|\alpha|_{*1} - |\alpha|_{*2,3} + |\alpha|_{*4} + 2 - |\alpha|_{\mathcal{A}} \\
&\leq 2|\alpha| - |\alpha|_+ - 2|\alpha|_\bullet + 1
\end{aligned}$$

By assumption, $|\alpha|_+ \geq 1$, any further binary operator pushes the right-hand side strictly below $2|\alpha'|$. Indeed, the only expression containing only one $+$ as binary operator, that reaches a conversion-ratio of 2, is $x_1^* + x_2^*$, which is of claimed structure. $\square$

Lemma 5.4. *A worst-case expression α has no subexpression of the form*

$$\phi = (\prod_i \sum_j \sigma_{ij}^*)^*$$

Proof. If $\phi \in sub(\alpha)$, ϵ-cycle elimination would occur upon expansion. By Cor. 5.1 then α would not be worst-case. $\square$

This allows us to provide a pretty detailed template of a worst-case expression:

Lemma 5.5. *Let α be worst-case. Then the structure of α is*

$$\alpha = \prod_{i=1}^{n} \sum_{j=1}^{k_i} \sigma_{ij}^*$$

where the σ_{ij} are literals.

Proof. By Prop. 5.2, a worst-case expression contains starred subexpressions, so fix some σ_{ij}^* which is by Lem. 5.3 operand to a sum. A maximal sum with stars is a factor, since it may not be starred itself and is already maximal. Further, σ_{ij} is necessarily a maximal product. If its operands were maximally starred sums, this would contradict Lem. 5.4, thus σ_{ij} is a product of literals. Then, σ_{ij} influences the conversion-ratio as given in Cor. 5.1 only by its length, which has to be minimized in order to maximize the ratio. Thus σ_{ij} is a literal. From Lem. 5.4 it also follows that α itself may not be starred. $\square$

It remains to analyze the influence of the number of summands (the k_i in Lem. 5.5) on conversion-ratio. This is done in the proof of our main

Theorem 5.2. *An expression α is worst-case, if its structure is*

$$\alpha = \prod_{i=1}^{n} \sum_{j=1}^{2+(i \bmod 2)} x_{ij}^* \quad \textit{where} \quad x_{ij} \in \mathcal{A} \cup \epsilon$$

Proof. Let α be of the general structure given in Lem. 5.5, the FA produced by a series of expansions from A_α^0 is sketched in Fig. 3. The sizes of these objects are

$$\begin{aligned} |\alpha| &= (n-1) + \sum_{i=1}^{n}(3k_i - 1) = 3\sum_{i=1}^{n} k_i - 1 \\ |A_\alpha| &= \sum_{i=1}^{n} 4k_i + n - 1 = 4\sum_{i=1}^{n} k_i + n - 1 \end{aligned}$$

thus the ratio is

$$\frac{|A_\alpha|}{|\alpha|} = \frac{4\sum k_i + n - 1}{3\sum k_i - 1} = 1 + \frac{\sum k_i + n}{3\sum k_i - 1}$$

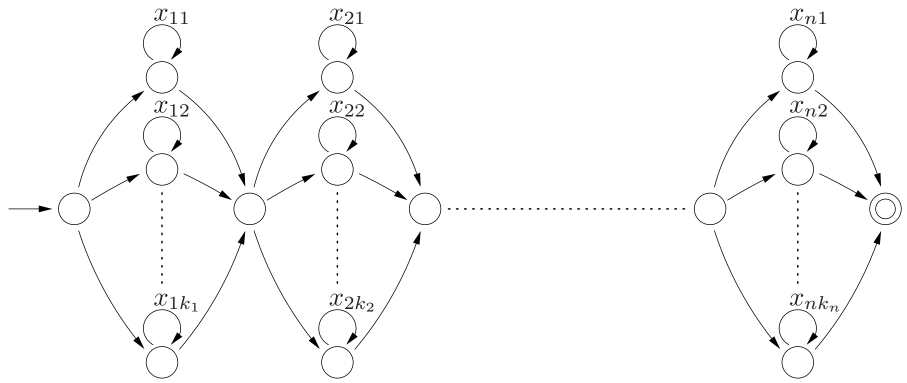


Figure 3: Automaton constructed from a worst-case expression (ϵ-labels are omitted).

The fraction on the right-hand side is maximized, if n is maximal with respect to $\sum k_i$, or equivalently, if $\sum k_i$ is minimal. Two restrictions result from prohibiting state-elimination, namely that $\forall i : k_i \geq 2$ and if k_i=2 then k_{i-1}>2 and k_{i+1}>2 (if they exist). Thus $\sum k_i$ is minimal, if k_i alternates between 2 and 3, i.e., $k_i = 2 + (i \bmod 2)$. □

Corollary 5.2. *The size of an automaton produced by our construction is bounded by* $\frac{22}{15}|\alpha| + 1$. *The construction is optimal.*

Proof. The value is reached by the expression given in Thm. 5.2, which was proven to give the maximal ratio of sizes. Since by Cor. 3.1 $\frac{22}{15}|\alpha| + 1$ is also a lower bound, the bound is tight, hence the construction is optimal. □

6 Conclusions & Remarks

We have given a construction for converting regular expressions into equivalent ϵNFAs. To our knowledge it is the only provably optimal construction so far. It should be mentioned that the generated automata differ from these constructed in [IY03] only by the effects of state-elimination. This element is crucial however, both for raising the lower bound as well as for upper bound analysis as we did.

Treatment of $\emptyset$ in expressions can easily be added to our algorithm by considering it a literal throughout the expansion/reduction-sequence and adding a final step: removing $\emptyset$-labeled transitions followed by running some reachability algorithm. The final step will reduce the size of the automaton, thus the bound is maintained even if $\emptyset$ does not count into the expressions' size. Since we consider $\emptyset$ as being of no practical relevance, it was omitted from formal treatment.

Maybe more interesting, Kleene-$+$ can be added by reformulating $*$-expansions, where additional ϵ-transitions need to be introduced. This yields smaller FAs than by applying the equivalence $\alpha^+ \equiv \alpha\alpha^*$ (which would double the number of elements introduced by α), yet it is not feasible with the given bound.

Finally we claim that $(\lhd \cup \rhd_q \cup \rhd_\gamma)$ is confluent as well, thus a series of combined expansions and eliminations converges to a unique FA. The proof involves more complex cases and is postponed to a future article.

References

[BS86] Gerard Berry and Ravi Sethi. From regular expressions to deterministic automata. *Theoretical Computer Science*, 48:117–126, 1986.

[EKSW05] Keith Ellul, Bryan Krawetz, Jeffrey Shallit, and Ming-Wei Wang. Regular expressions: new results and open problems. *Journal of Automata, Languages and Combinatorics*, 10(4):407–437, 2005.

[Glu61] Victor Michailowitsch Glushkov. The abstract theory of automata. *Russian Mathematical Surveys*, 16:1–53, 1961.

[Hue80] Gérard Huet. Confluent reductions: Abstract properties and applications to term rewriting systems. *Journal of the ACM*, 27(4):797–821, 1980.

[IY03] Lucian Ilie and Sheng Yu. Follow automata. *Information and Computation*, (186):140–162, 2003.

[Kle65] Stephen Cole Kleene. *Representation of Events in Nerve Nets and Finite Automata*, pages 3–41. Annals of Mathematics Studies. 1965.

[MY60] Robert McNaughton and Hisao Yamada. Regular expressions and state graphs for automata. *IRE Transactions on Electronic Computers*, 9(1):39–47, 1960.

[OF61] Gene Ott and Neil H. Feinstein. Design of sequential machines from their regular expressions. *Journal of the ACM*, 8(4):585–600, 1961.

[SSS88] Seppo Sippu and Eljas Soisalin-Soininen. *Parsing Theory*. EATCS Monographs on Theoretical Computer Science. Springer, 1988.

[Tho68] Ken Thompson. Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422, 1968.

[Wat94] Bruce W. Watson. A taxonomy of finite automata construction algorithms. Technical Report Computing Science Note 93/43, Eindhoven University of Technology, may 1994.

[Woo87] Derick Wood. *Theory of Computation*. John Wiley & Sons, Inc., 1987.

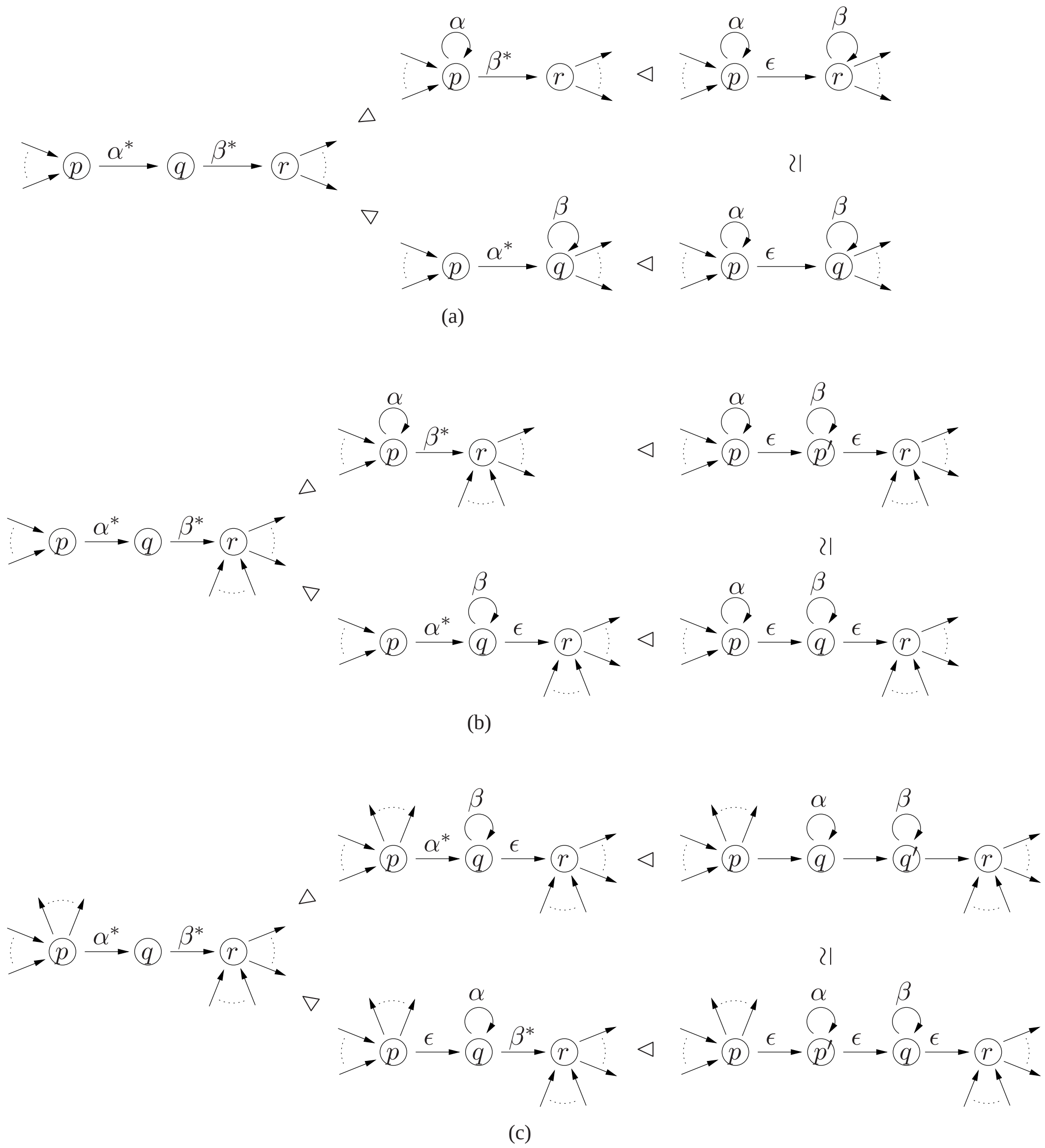


Figure 4: Examples for confluence of expanding consecutive starred transitions. Isomorphism is denoted by $\simeq$.