Regular Expression

Beulah A.
AP/CSE

Introduction

- Regular expressions describe regular languages
- ie the language accepted by a finite automata are easily described by regular expression.
- Many programming languages provide regular expression capabilities,
  - Built-in → Perl, JavaScript, Ruby, AWK, Tcl,
  - Standard library →.NET, Java, Python C++
- REs are widely supported in programming languages, text processing programs (particular lexers, lex, yacc), advanced text editors

Introduction

- Let Σ be a finite set of symbols.
- Let L_1, L_2 be set of strings in Σ^*.
- The concatenation of L_1 and L_2 denoted by $L_1 L_2$ is the set of all strings of the form *xy*, where $x \in L_1$ and $y \in L_2$.
- $L^0 = \{\varepsilon\}$
- $L^i = LL^{i-1}$ for $i \geq 1$.

Introduction

- Kleene Closure

$$L^* = \bigcup_{i=0}^{\infty} L^i = L^0 \cup L^1 \cup L^2 \cup \ldots$$

- Positive Closure

$$L^+ = \bigcup_{i=1}^{\infty} L^i = L^1 \cup L^2 \cup \ldots$$

Example

Let $L_1 = \{10, 01\}$, $L_2 = \{11, 00\}$

Then $L_1L_2 = \{1011, 1000, 0111, 0100\}$

Let $L = \{10, 11\}$

Then $L^* = L_0 \cup L_1 \cup L_2 \cup \ldots\ldots$

$= \{\varepsilon\} \cup \{10, 11\} \cup \{1011, 1010, 1110, 1111\} \cup \ldots.$

$= \{\varepsilon, 10, 11, 1011, 1010, 1110, 1111, \ldots\ldots.\}$

Operators of RE

$* \rightarrow L^*$

$. \rightarrow L_1 . L_2 , L_1L_2$

$/ \rightarrow L_1 \cup L_2$

Definition of Regular Expression

- Let Σ be an alphabet. The regular expressions over Σ and the sets that they denote are defined recursively as follows:

1. φ is a regular expression and denotes the empty set {}.
2. ε is a regular expression and denotes the set {ε}
3. For each a ∈ Σ, 'a' is a regular expression and denotes the set {a}.
4. If r and s are regular expressions denoting the languages R and S respectively then (r + s), (rs), (r)* are regular expressions that denotes the sets R∪S, RS and R* respectively.

Precedence of RE operators

* → higher precedence

.

/ → Lower precedence

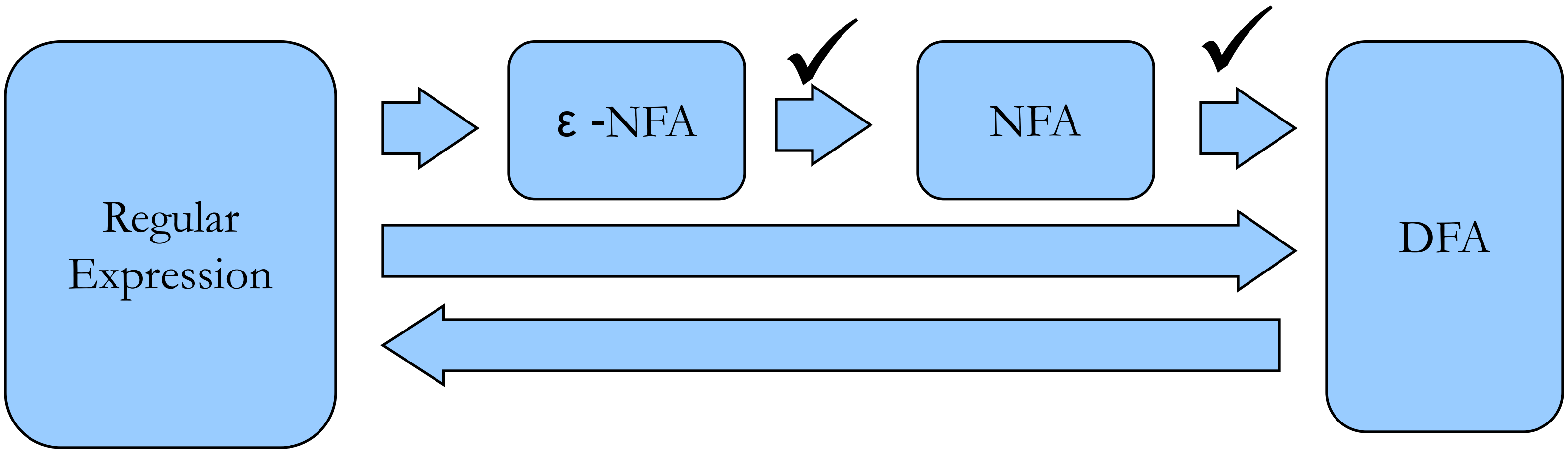
Example

- (0/1)* = {ε,0,1,00, 01, 10, 11.....} = (0+1)*
  (i.e.) all strings of 0 and 1
- 01* = {0, 01, 011, 0111, ...}
- 0* = {ε, 0, 00, 000, ...}
- 1(1)* = {1, 11, 111, 1111, ...) = 1^+

Identities for Regular Expressions

I1	$\varphi + R = R$
I2	$\varphi R = R\varphi = \varphi$
I3	$\lambda R = R\lambda = R$
I4	$\lambda^* = \lambda$
I5	$R + R = R$
I6	$R^*R^* = R^*$
I7	$RR^* = R^*R$
I8	$(R^*)^* = R^*$
I9	$\lambda + RR^* = R^* = \lambda + R^*R$
I10	$(PQ)^*P = P(QP)^*$
I11	$(P + Q)^* = (P^*Q^*)^* = (P^* + Q^*)^*$
I12	$(P + Q)R = PR + QR$ and $R(P + Q) = RP + RQ$

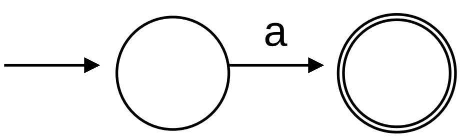
Road map



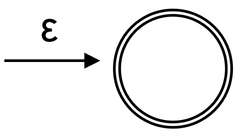
Thompson's Construction

- Basis

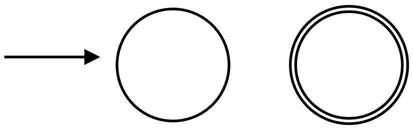
R=a



R=ε

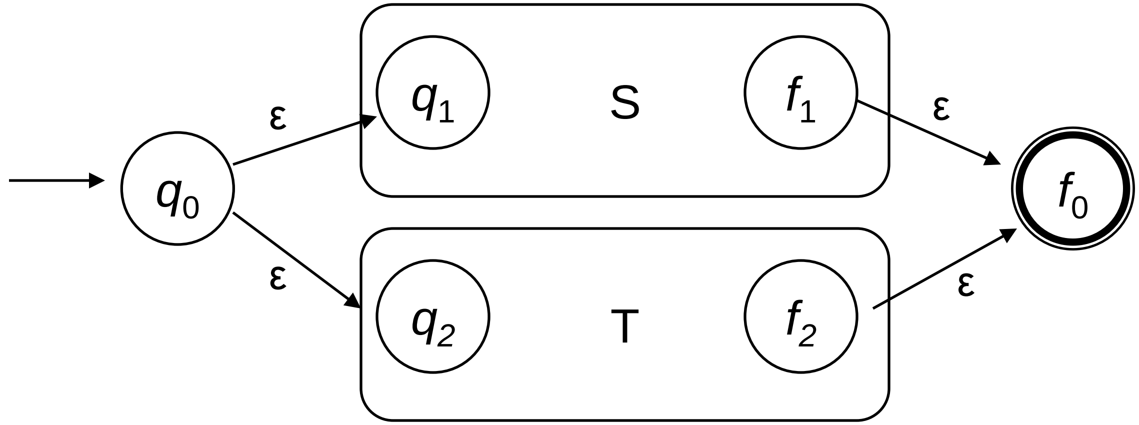


R= φ

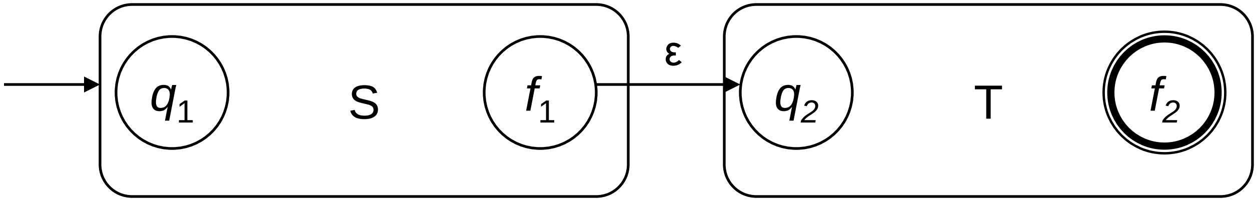


Thompson's Construction

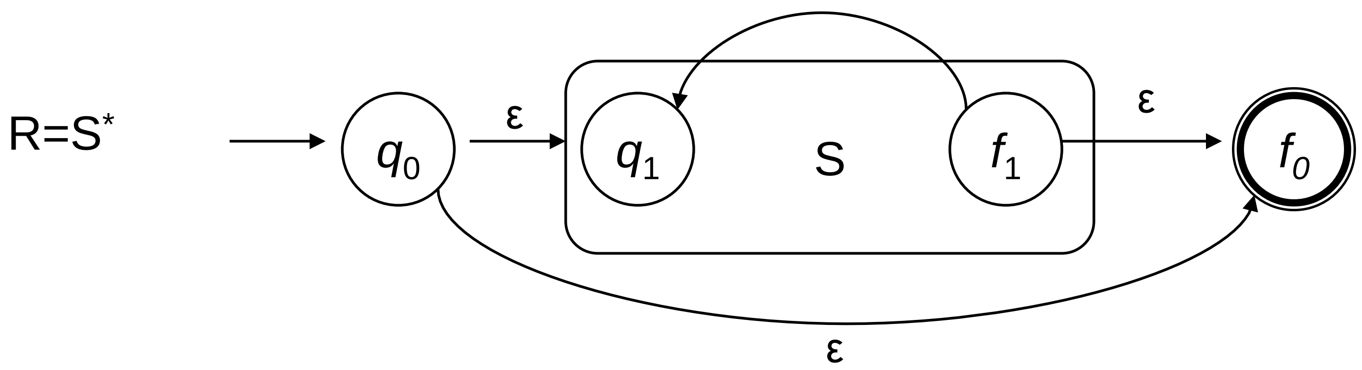
R=S+T



R=ST



Thompson's Construction



Theorem

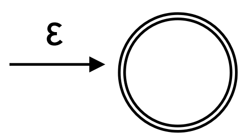
For every regular expression *r* there exists a NFA with ε-transitions that accepts L(*r*)

- Proof
  - **Basis step (Zero operators)**

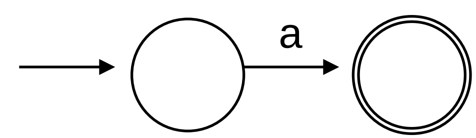
Suppose *r* is ε, φ or *a* for some *a* ∈ ∑.

Then the equivalent NFA's are:

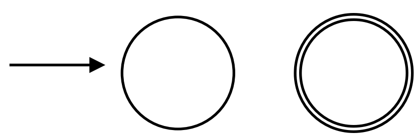
R=ε



R=a



R= φ



Induction Case i

- $r = r_1 + r_2$
- $M_1 = (Q_1, \sum_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = L(r_1)$
- $M_2 = (Q_2, \sum_2, \delta_2, q_2 \{f_2\})$ $L(M_2) = L(r_2)$.
- Assume Q_1 and Q_2 are disjoint.
- Let q_0, f_0 be a new initial and final state respectively.

Case i

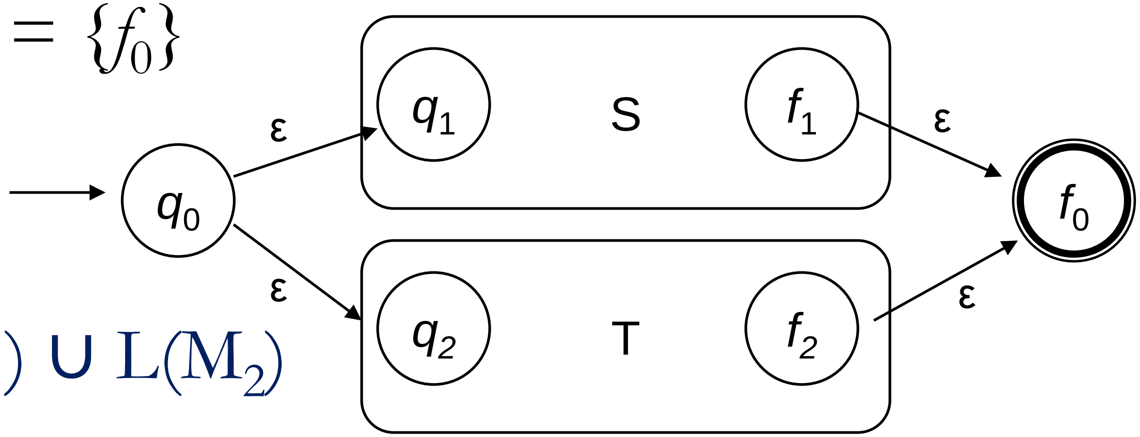
- $M = (Q_1 \cup Q_2 \cup \{q_0, f_0\}, \Sigma_1 \cup \Sigma_2, \delta, q_0, \{f_0\})$ where δ is defined by

$\delta(q_0, \varepsilon) = \{q_1, q_2\}$

$\delta(q, a) = \delta_1(q, a)$ if $q \in Q_1 - \{f_1\}$, $a \in \Sigma_1 \cup \{\varepsilon\}$

$\delta(q, a) = \delta_2(q, a)$ if $q \in Q_2 - \{f_2\}$, $a \in \Sigma_2 \cup \{\varepsilon\}$

$\delta_1(f_1, \varepsilon) = \delta_2(f_2, \varepsilon) = \{f_0\}$



- $L(M) = L(M_1) \cup L(M_2)$

Case ii

- $r = r_1 \ . \ r_2$
- $M_1 = (Q_1, \Sigma_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = L(r_1)$
- $M_2 = (Q_2, \Sigma_2, \delta_2, q_2 \{f_2\})$ $L(M_2) = L(r_2)$
- $M = (Q_1 \cup Q_2, \Sigma_1 \cup \Sigma_2, \delta, \{q_1\}, \{f_2\})$, where δ is given by:

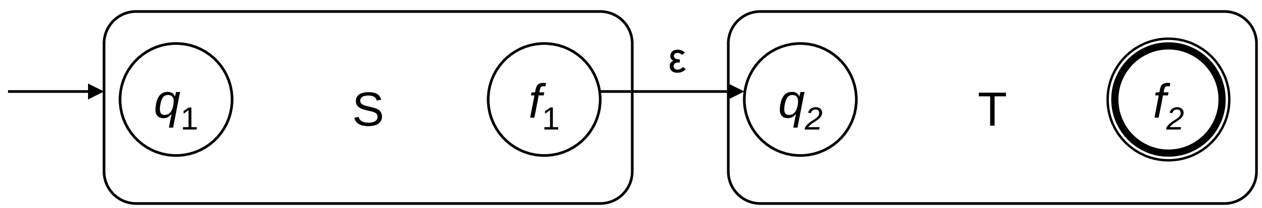
$\delta(q,a) = \delta_1(q,a)$ for q in $Q_1 - \{f_1\}$ and a in $\Sigma_1 \cup \{\varepsilon\}$

$\delta(f_1,\varepsilon) = \{q_2\}$

$\delta(q,a) = \delta_2(q,a)$ for q in Q_2 and a in $\Sigma_2 \cup \{\varepsilon\}$

Case ii

- $L(M) = \{xy \mid x$ is in $L(M_1)$ and y is in $L(M_2)\}$
- $L(M) = L(M_1) \,.\, L(M_2)$.

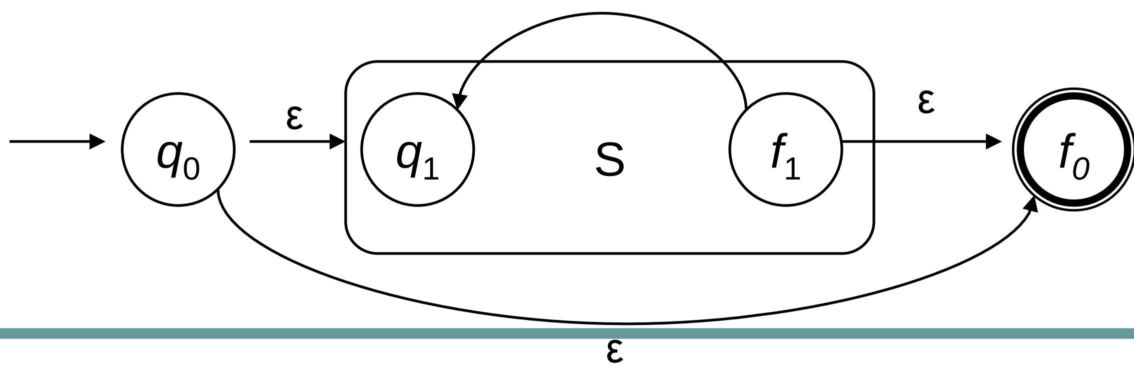


Case iii

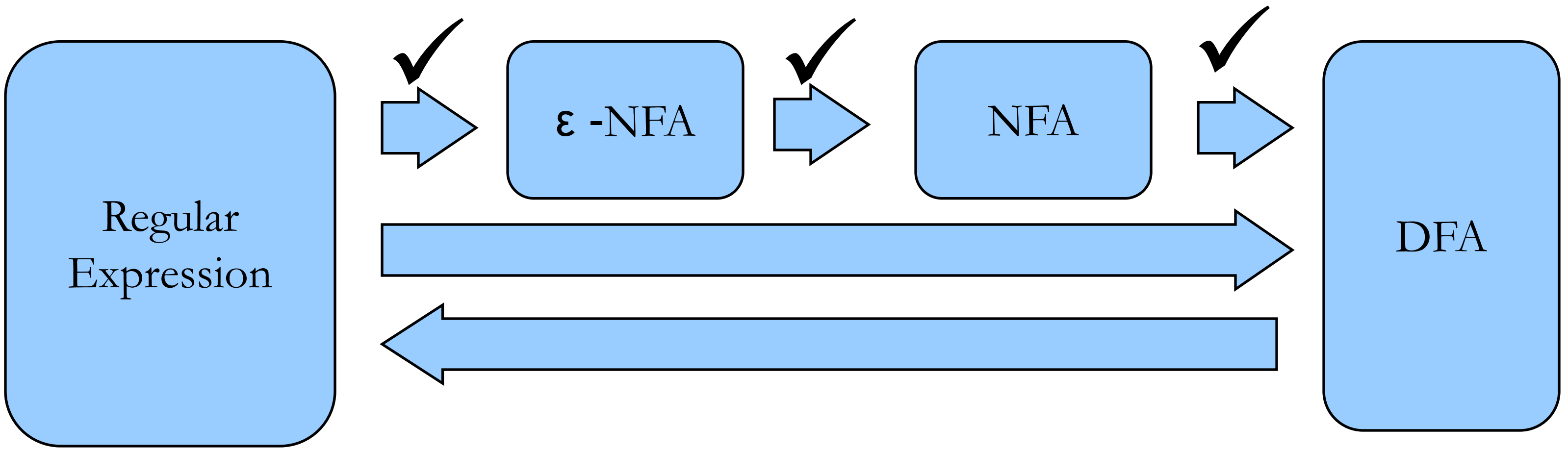
- $r = r_1^*$
- $M_1 = (Q_1, \sum_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = r_1$
- $M = (Q_1 \cup \{q_0, f_0\}, \sum_1, \delta, q_0, \{f_0\})$, where δ is given by:

$\delta(q, \varepsilon) = \delta(f_1, \varepsilon) = \{q_1, f_0\}$

$\delta(q, a) = \delta_1(q, a)$ for q in $Q_1 - \{f_1\}$ and a in $\sum_1 \cup \{\varepsilon\}$



Road map



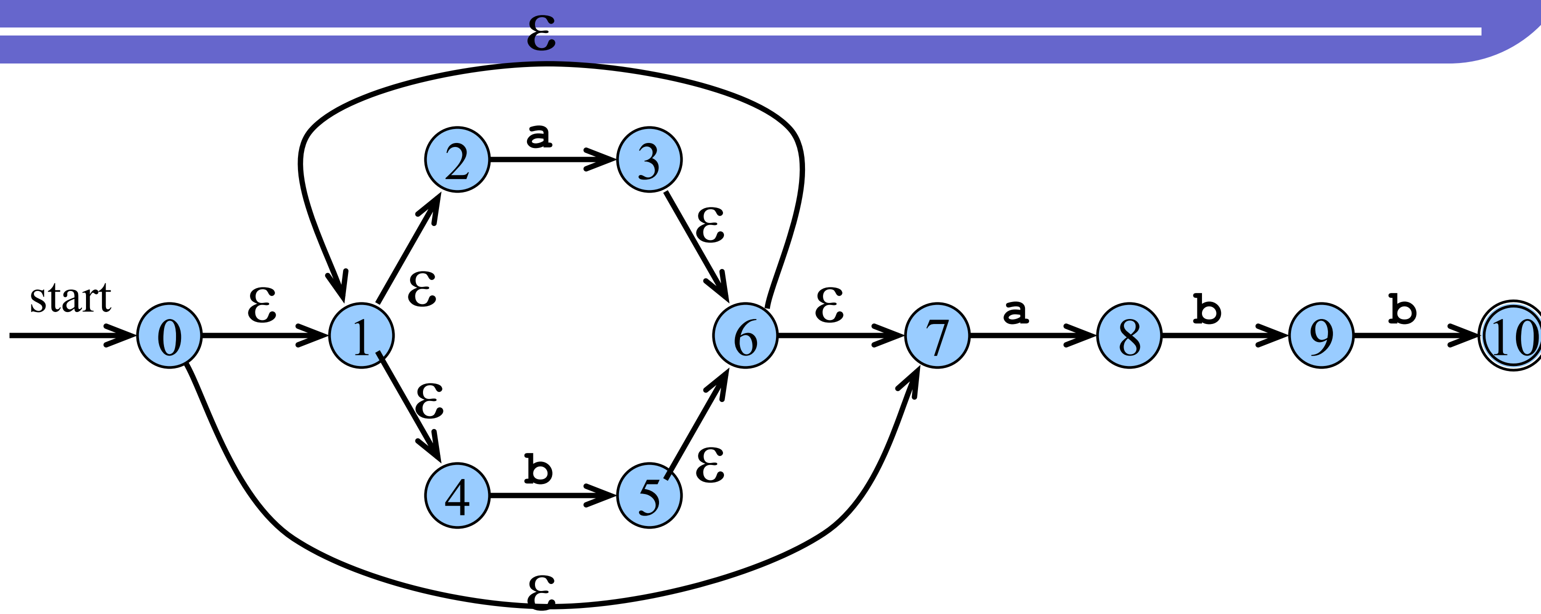
Conversion of ε-NFA to DFA

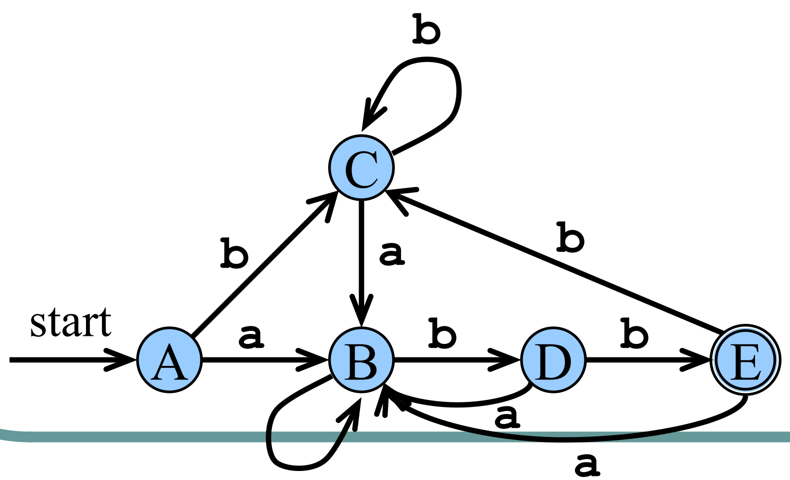
1. Find the ε-CLOSURE of the state q_0 from the constructed ε-NFA (i.e) from state q_0, ε transition to other states are identified as well as ε transitions from other states are also identified and combined as one set (new state).

Conversion of ε-NFA to DFA

2. Perform the following steps until there are no more new states as been constructed.
   i. Find the transition of the given regular expression symbols over ∑ from the new state (i.e) move (new state, symbol)
   ii. Find the ε-CLOSURE of move (new state, symbol).

Example





Dstates

A = {0,1,2,4,7}

B = {1,2,3,4,6,7,8}

C = {1,2,4,5,6,7}

D = {1,2,4,5,6,7,9}

E = {1,2,4,5,6,7,10}

Summary

- Definition of RE
- Precedence, identities, properties of RE.
- Thomson's construction to convert RE to NFA and then to DFA

Test Your Knowledge

- Which of the following does not represents the given language?

  Language: {0,01}

  a) 0+01

  b) {0} U {01}

  c) {0} U {0}{1}

  d) {0} ^ {01}

Test Your Knowledge

- Regular Expression R and the language it describes can be represented as:
  a) R, R(L)
  b) L(R), R(L)
  c) R, L(R)
  d) All of the mentioned

Reference

- Hopcroft J.E., Motwani R. and Ullman J.D, "Introduction to Automata Theory, Languages and Computations", Second Edition, Pearson Education, 2008