CMPE 322/327 - Theory of Computation

Week 4: Pattern Matching & Regular Expressions

Burak Ekici

March 14-18, 2022

Outline

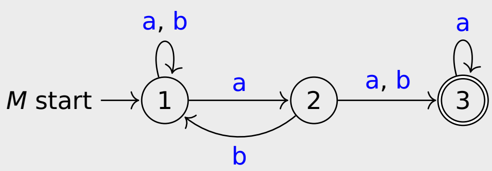
1. A Quick Recap
2. Pattern Matching
3. Regular Expressions
4. Homomorphisms

Definitions

- nondeterministic finite automaton (NFA) is quintuple $N = (Q, \Sigma, \Delta, s, F)$ with
  1. Q : finite set of states
  2. Σ : input alphabet
  3. $\Delta : Q \times \Sigma \to 2^Q$: transition function
  4. $S \subseteq Q$: start state
  5. $F \subseteq Q$: final (accept) states

Example

$N = (Q, \Sigma, \Delta, S, F)$



1. $Q := \{1, 2, 3\}$
2. $\Sigma := \{a, b\}$
3. $\Delta : Q \times \Sigma \to 2^Q$
4. $S := \{1\}$
5. $F := \{3\}$

Δ	a	b
1	$\{1, 2\}$	$\{1\}$
2	$\{3\}$	$\{1, 3\}$
3	$\{3\}$	$\emptyset$

Definitions

- nondeterministic finite automaton (NFA) is quintuple $N = (Q, \Sigma, \Delta, s, F)$ with
  1. Q : finite set of states
  2. Σ : input alphabet
  3. $\Delta : Q \times \Sigma \to 2^Q$: transition function
  4. $S \subseteq Q$: set of start states
  5. $F \subseteq Q$: final (accept) states
- $\widehat{\Delta} : 2^Q \times \Sigma^* \to 2^Q$ is inductively defined by

$$\widehat{\Delta}(A, \varepsilon) = A \qquad\qquad \widehat{\Delta}(A, xa) = \bigcup_{q \in \widehat{\Delta}(A,x)} \Delta(q, a)$$

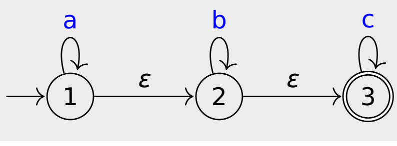
- string $x \in \Sigma^*$ is accepted by N if $\widehat{\Delta}(S, x) \cap F \neq \varnothing$

Definitions

- NFA with ε-transitions (NFA$_\varepsilon$) is sextuple $N = (Q, \Sigma, \varepsilon, \Delta, S, F)$ such that
  1. $\varepsilon \notin \Sigma$
  2. $N_\varepsilon = (Q, \Sigma \cup \{\varepsilon\}, \Delta, S, F)$ is NFA over alphabet $\Sigma \cup \{\varepsilon\}$
- ε-closure of set $A \subseteq Q$ is defined as $C_\varepsilon(A) = \bigcup \{ \widehat{\Delta}_{N_\varepsilon}(A, x) \mid x \in \{\varepsilon\}^* \}$
- $\widehat{\Delta}_N : 2^Q \times \Sigma^* \to 2^Q$ is inductively defined by

$$\widehat{\Delta}_N(A, \varepsilon) = C_\varepsilon(A) \qquad\qquad \widehat{\Delta}_N(A, xa) = \bigcup \{ C_\varepsilon(\Delta(q, a)) \mid q \in \widehat{\Delta}_N(A, x) \}$$

Example



$$C_\varepsilon(\{1\}) = \{1, 2, 3\}$$
$$\widehat{\Delta}(\{1\}, b) = \{2, 3\}$$

Theorem

every set accepted by NFA is regular

Theorem

every set accepted by NFA_ϵ is regular

Theorem

regular sets are effectively closed under concatenation

Theorem

regular sets are effectively closed under asterate

Outline

1. A Quick Recap
2. Pattern Matching
3. Regular Expressions
4. Homomorphisms

Pattern matching is important for

- lexical analysis of programs
- scripting languages (Perl, Ruby)
- search engines (Google Code Search)
- DNA analysis

Applications of Regular expressions: `grep`

- `grep foo file` returns lines in file containing pattern foo
- basis for more powerful tools like `awk`, `sed`, `perl`

Some Patterns

`^`	matches beginning of line	`.`	matches any character
`$`	matches end of line	`[abc]`	matches a or b or c
`c`	matches character c	`[a-zA-Z]`	matches any letter

Example

`grep "0" file`	returns lines containing 0
`grep "0$" file`	returns lines ending with 0
`grep "b.g" file`	returns lines containing e.g. bag, big, bug, buggy

Pattern matching is important for

- lexical analysis of programs
- scripting languages (Perl, Ruby)
- search engines (Google Code Search)
- DNA analysis

Definitions

- pattern is string α that represents set of strings $L(\alpha) \subseteq \Sigma^*$
-

atomic pattern α	$L(\alpha)$
$\mathbf{a} \in \Sigma$	$\{a\}$
$\boldsymbol{\epsilon}$	$\{\epsilon\}$
$\boldsymbol{\varnothing}$	$\varnothing$
$\#$	Σ
$@$	Σ^*

compound pattern α	$L(\alpha)$
$\beta + \gamma$	$L(\beta) \cup L(\gamma)$
$\beta \cap \gamma$	$L(\beta) \cap L(\gamma)$
$\beta\gamma$	$L(\beta)L(\gamma)$
β^*	$L(\beta)^*$
β^+	$L(\beta)^+$
$\sim \beta$	$\sim L(\beta) = \Sigma^* - L(\beta)$

- string $x \in \Sigma^*$ matches pattern α if $x \in L(\alpha)$

Example

pattern	matched string
$@a@a@a@$	strings containing at least 3 occurrences of a
$@a@b@$	strings containing a followed later by b
$\# \cap \sim a$	single letters except a
$(\# \cap \sim a)^*$	strings without a

Questions

- how difficult is pattern matching?
- is pattern equivalence $(L(\alpha) = L(\beta))$ decidable?
- which operators are redundant?

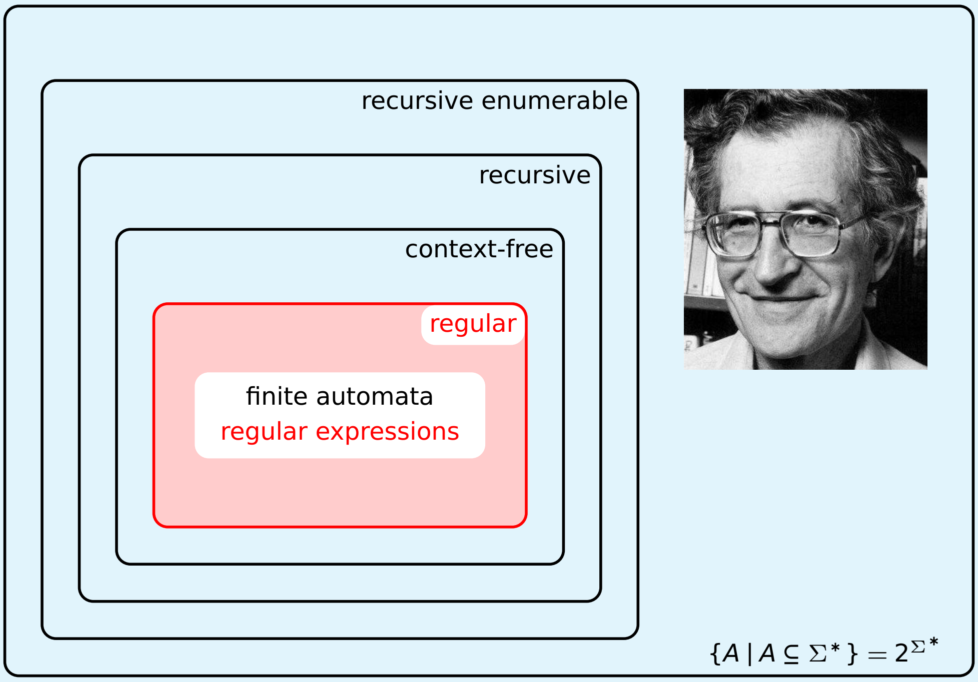
$$\begin{array}{lcll} \boldsymbol{\varepsilon} & \equiv & \sim(\#@) \equiv \boldsymbol{\varnothing}^* & \\ @ & \equiv & \#^* & \\ \alpha^+ & \equiv & \alpha\alpha^* & \\ \# & \equiv & a_1 \ldots a_n & \text{if } \Sigma = \{a_1 \ldots a_n\} \\ \alpha \cap \beta & \equiv & \sim(\sim\alpha + \sim\beta) & \\ \sim\alpha & \equiv & ? & \end{array}$$

Notation

$\alpha \equiv \beta$ if $L(\alpha) = L(\beta)$

Outline

1 A Quick Recap
2 Pattern Matching
3 Regular Expressions
4 Homomorphisms



Definition

regular expressions are restricted patterns which use only

$$\boldsymbol{a} \in \Sigma \quad \boldsymbol{\varepsilon} \quad \boldsymbol{\varnothing} \quad \alpha + \beta \quad \alpha^* \quad \alpha\beta$$

Theorem

finite automata, patterns, and regular expressions are equivalent:

for all $A \subseteq \Sigma^*$ 1 A is regular

$\iff$ 2 $A = L(\alpha)$ for some pattern α

$\iff$ 3 $A = L(\alpha)$ for some regular expression α

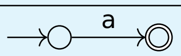
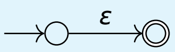
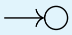
Proof.

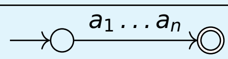
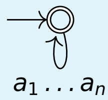
3	$\Longrightarrow$	2	trivial	(every regular expression is a pattern)
2	$\Longrightarrow$	1	induction on α	(see slides #17 − 18)
1	$\Longrightarrow$	3		(see slide #19)

Proof. (2 $\implies$ 1)

statement: for any pattern α, $L(\alpha)$ is regular
induction on pattern α

1. atomic pattern $\Sigma = \{a_1, \ldots, a_n\}$

α	$L(\alpha)$	finite automaton
$\boldsymbol{a} \in \Sigma$	$\{a\}$	
$\boldsymbol{\varepsilon}$	$\{\varepsilon\}$	
$\boldsymbol{\varnothing}$	$\varnothing$	

α	$L(\alpha)$	finite automaton
#	Σ	
@	Σ^*	

Proof. (2 $\implies$ 1)

statement: for any pattern α, $L(\alpha)$ is regular
induction on pattern α

2. compound patterns

α	$L(\alpha)$
$\beta + \gamma$	$L(\beta) \cup L(\gamma)$
$\beta \cap \gamma$	$L(\beta) \cap L(\gamma)$
$\beta\gamma$	$L(\beta)L(\gamma)$

α	$L(\alpha)$
β^*	$L(\beta)^*$
β^+	$L(\beta)^+$
$\sim\beta$	$\sim L(\beta)$

$L(\beta)$ and $L(\gamma)$ are regular according to induction hypothesis
hence $L(\alpha)$ is regular according to closure properties of regular sets

Proof. (1 $\Longrightarrow$ 3 – An idea)

given NFA$_\varepsilon$ $N_\varepsilon = (Q, \Sigma, \varepsilon, \Delta, S, F)$

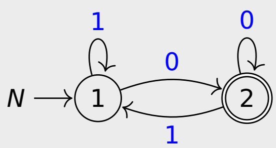
$\forall Y \subseteq Q \quad \forall u, v \in Q$ construct regular expression α^Y_{uv} such that

$$x \in L(\alpha^Y_{uv}) \iff \boxed{\begin{array}{l}\exists \text{ a path from } u \text{ to } v \text{ labeled } x \; (v \in \widehat{\Delta}(\{u\}, x)) \\ \text{such that all intermediate states belong to } Y\end{array}}$$

Definitions

- $\alpha^{\emptyset}_{uv} := \begin{cases} \boldsymbol{a_1} + \ldots + \boldsymbol{a_k} & \text{if } u \neq v \text{ and } k > 0 \\ \boldsymbol{\emptyset} & \text{if } u \neq v \text{ and } k = 0 \\ \boldsymbol{a_1} + \ldots + \boldsymbol{a_k} + \boldsymbol{\varepsilon} & \text{if } u = v \text{ and } k > 0 \\ \boldsymbol{\varepsilon} & \text{if } u = v \text{ and } k = 0 \end{cases}$ $\quad \{a_1, \ldots, a_k\} := \{a \in \Sigma \cup \{\varepsilon\} \mid v \in \Delta(u, a)\}$
- $\alpha^Y_{uv} := \alpha^{Y-\{q\}}_{uv} + \alpha^{Y-\{q\}}_{uq} (\alpha^{Y-\{q\}}_{qq})^* \alpha^{Y-\{q\}}_{qv}$ for some fixed $q \in Y$

Example



$L(N) = L(\alpha)$ with

$\alpha = \alpha^{\{1,2\}}_{12} = \alpha^{\{1\}}_{12} + \alpha^{\{1\}}_{12} (\alpha^{\{1\}}_{22})^* \alpha^{\{1\}}_{22}$ (q = 2)

$\alpha^{\{1\}}_{12} = \alpha^{\emptyset}_{12} + \alpha^{\emptyset}_{11} (\alpha^{\emptyset}_{11})^* \alpha^{\emptyset}_{12} = 0 + (1 + \boldsymbol{\varepsilon})(1 + \boldsymbol{\varepsilon})^* 0$

$\alpha^{\{1\}}_{22} = \alpha^{\emptyset}_{22} + \alpha^{\emptyset}_{21} (\alpha^{\emptyset}_{11})^* \alpha^{\emptyset}_{12} = (0 + \boldsymbol{\varepsilon}) + 1(1 + \boldsymbol{\varepsilon})^* 0$

$\alpha^{\emptyset}_{12} = 0 \quad \alpha^{\emptyset}_{11} = 1 + \boldsymbol{\varepsilon} \quad \alpha^{\emptyset}_{22} = 0 + \boldsymbol{\varepsilon} \quad \alpha^{\emptyset}_{21} = 1$

$\alpha = (0 + (1 + \boldsymbol{\varepsilon})(1 + \boldsymbol{\varepsilon})^* 0) + (0 + (1 + \boldsymbol{\varepsilon})(1 + \boldsymbol{\varepsilon})^* 0)((0 + \boldsymbol{\varepsilon}) + 1(1 + \boldsymbol{\varepsilon})^* 0)^* ((0 + \boldsymbol{\varepsilon}) + 1(1 + \boldsymbol{\varepsilon})^* 0)$

$\equiv (0 + 1)^* 0$

Outline

1 A Quick Recap

2 Pattern Matching

3 Regular Expressions

4 Homomorphisms

Theorem

regular sets are effectively closed under homomorphic image and preimage

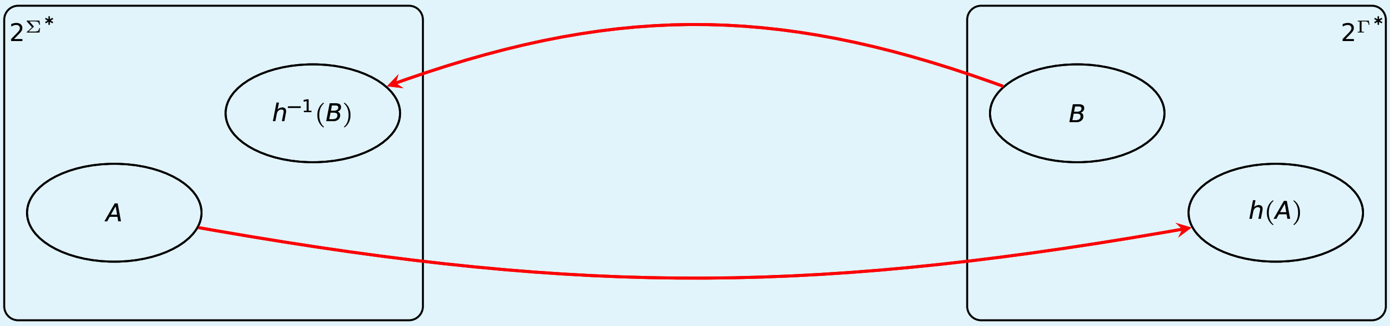
Definitions

- homomorphism is mapping $h\colon \Sigma^* \to \Gamma^*$ such that

$$h(\epsilon) = \epsilon \qquad\qquad h(xy) = h(x)h(y)$$

  so homomorphism is completely determined by its effect on Σ

- if $A \subseteq \Sigma^*$ then $h(A) = \{h(x) \mid x \in A\} \subseteq \Gamma^*$ "image of A under h"
  if $B \subseteq \Gamma^*$ then $h^{-1}(B) = \{x \mid h(x) \in B\} \subseteq \Sigma^*$ "preimage of B under h"



- homomorphism $h\colon \Sigma^* \to \Gamma^*$
- $h^{-1}(h(A)) \supseteq A$
- $h(h^{-1}(B)) \subseteq B$

Example

$\Sigma = \Gamma = \{0,1\}$ $\quad h(0) = 11 \quad h(1) = 1 \quad A = \{0\}\Sigma = \Gamma = \{0,1\}$
$h(0) = 11 \quad h(1) = 1 \quad A = B = \{0\}$

- $h^{-1}(h(A)) = h^{-1}(\{11\}) = \{0, 11\} \supset A$
- $h(h^{-1}(B)) = h(\varnothing) = \varnothing \subset B$

Lemma

$A \subseteq \{0,1\}^*$ is regular $\quad\Longrightarrow\quad$ $\{xy \mid x1y \in A\}$ is regular

Proof.

- $\Sigma = \{0,1\}$ and $\Gamma = \{0,1,2\}$
- define homomorphisms $h, i\colon \Gamma^* \to \Sigma^*$ by

$$h(0) = 0 \quad h(1) = h(2) = 1 \quad i(0) = 0 \quad i(1) = 1 \quad i(2) = \varepsilon$$

- $h^{-1}(A) = \{x \mid h(x) \in A\}$
- $h^{-1}(A) \cap L((0+1)^*2(0+1)^*) = \{x2y \mid x1y \in A\}$
- $\{xy \mid x1y \in A\} = i(h^{-1}(A) \cap L((0+1)^*2(0+1)^*))$ is regular

Theorem

regular sets are effectively closed under homomorphic image and preimage

Proof.

- DFA $M = (Q, \Gamma, \delta, s, F)$
- homomorphism $h \colon \Sigma^* \to \Gamma^*$
- $h^{-1}(L(M)) = L(M')$ for DFA $M' = (Q, \Sigma, \delta', s, F)$ with $\delta'(q, a) := \widehat{\delta}(q, h(a))$
- claim: $\widehat{\delta'}(q, x) = \widehat{\delta}(q, h(x))$ $\quad \forall x \in \Sigma^* \ \forall q \in Q$
  proof of claim: induction on $|x|$ (see next slide)

A Quick Recap | Pattern Matching | Regular Expressions | Homomorphisms

proof of the claim

claim: $\widehat{\delta'}(q, x) = \widehat{\delta}(q, h(x)) \quad \forall x \in \Sigma^* \ \forall q \in Q$

- base case: $|x| = 0$ thus $x = \varepsilon$

$$\widehat{\delta'}(q, \varepsilon) = q = \widehat{\delta}(q, h(\varepsilon))$$

- step case: $|x| > 0$ thus $x = ya$ s.t. $|y| = |x| - 1$ with IH : $\widehat{\delta'}(q, y) = \widehat{\delta}(q, h(y))$

$$\begin{aligned}
\widehat{\delta'}(q, ya) &= \delta'(\widehat{\delta'}(q, y), a) && \text{(by definition of } \widehat{\delta'}\text{)} \\
&= \delta'(\widehat{\delta}(q, h(y)), a) && \text{(by induction hypothesis IH)} \\
&= \widehat{\delta}(\widehat{\delta}(q, h(y)), h(a)) && \text{(by definition of } \delta'\text{)} \\
&= \widehat{\delta}(q, h(y)h(a)) && \text{(by distributivity of } \widehat{\delta} \text{ – w3.pdf, slide 10)} \\
&= \widehat{\delta}(q, h(ya)) && \text{(by definition of homomorphism)} \\
&= \widehat{\delta}(q, h(x))
\end{aligned}$$

$\square$

Proof. (closedness under complement homomorphic preimage)

statement: $L(M') = h^{-1}(L(M))$

$$\begin{aligned}
\forall x \in \Sigma^*,\ x \in L(M') &\iff \widehat{\delta'}(s, x) \in F && \text{(by definition of acceptance)}\\
&\iff \widehat{\delta}(s, h(x)) \in F && \text{(by claim proven in slide 26)}\\
&\iff h(x) \in L(M) && \text{(by definition of acceptance)}\\
&\iff x \in h^{-1}(L(M)) && \text{(by definition of homomorphic preimage)}
\end{aligned}$$

□

Example

This page has too many overlays. Please refer to the original slides (w4.pdf) to monitor the whole content.

Theorem

regular sets are effectively closed under homomorphic image and preimage

Proof.

- regular expression α over Σ
- homomorphism $h\colon \Sigma^* \to \Gamma^*$
- $h(L(\alpha)) = L(\alpha')$ for regular expression α' defined inductively:

$$
\begin{array}{lcll}
\mathbf{a}' & = & h(\mathbf{a}) & \text{for } \mathbf{a} \in \Sigma \\
\boldsymbol{\epsilon}' & = & \boldsymbol{\epsilon} & \\
\boldsymbol{\varnothing}' & = & \boldsymbol{\varnothing} &
\end{array}
\qquad
\begin{array}{lcl}
(\beta+\gamma)' & = & \beta'+\gamma' \\
(\beta\gamma)' & = & \beta'\gamma' \\
(\beta^*)' & = & (\beta')^*
\end{array}
$$

Definitions

- Hamming distance $H(x,y)$ is number of places where bit strings x and y differ (if $|x| \neq |y|$ then $H(x,y) = \infty$)
- $N_k(A) := \{x \in \{0,1\}^* \mid H(x,y) \leqslant k \text{ for some } y \in A\}$

Lemma

$A \subseteq \{0,1\}^*$ is regular $\implies$ $\forall k \in \mathbb{N}$, $N_k(A)$ is regular

Proof.

$$
\begin{array}{lcll}
D_k & = & \{x \in (\{0,1\} \times \{0,1\})^* \mid x \text{ contains at most } k \text{ pairs } (0,1) \text{ or } (1,0)\} & \text{is regular} \\
 & = & \{x \in (\{0,1\} \times \{0,1\})^* \mid H(\mathit{fst}(x), \mathit{snd}(x)) \leqslant k\} & \\
N_k(A) & = & \mathit{fst}(\mathit{snd}^{-1}(A) \cap D_k) &
\end{array}
$$

Example

This page has too many overlays. Please refer to the original slides (w4.pdf) to monitor the whole content.

Thanks! & Questions?