4–5. Regularitat

Teoria de la Computació

FIB

Antoni Lozano

Q2 2024–2025

Regularitat

Concepte

Les expressions regulars són notacions per descriure patrons de text que inclouen les operacions de

- concatenació,
- reunió i
- estrella de Kleene.

Exemple 1: Paraules que acaben en *ab*

$$(a+b)^*ab$$

Exemple 2: Mots sobre $\{0,1\}$ amb un nombre parell de zeros

$$(01^*0+1)^*$$

Concepte

Les expressions regulars

- van ser inventades per Stephen Kleene a mitjans dels 1950s com una notació per als autòmats finits
- representen llenguatges equivalents als dels DFAs (regulars) mitjançant tres operacions per les quals els regulars són tancats
- es van fer servir ben aviat dins d'editors de text, en l'eina grep de Unix o en llenguatges com Pearl

Definició

Definició: Expressió regular

Sigui Σ un alfabet. Diem que r és una expressió regular (*e.r.*) si r és

1. $\emptyset$
2. λ
3. a, per a un símbol $a \in \Sigma$
4. $(r_1 + r_2)$, on r_1 i r_2 són expressions regulars
5. $(r_1 \cdot r_2)$, on r_1 i r_2 són expressions regulars
6. (r_1^*), on r_1 és una expressió regular

Definició

Exemple: e.r. simplificada

L'expressió regular vista abans

$$(01^*0 + 1)^*$$

es construeix formalment com

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

Simplificacions

- Per estalviar parèntesis, considerem precedències en els operadors. De més a menys prioritat: $*$, $\cdot$, $+$
- S'el·lideix 'operador $\cdot$
- Si r és una e.r., considerem
  - r^+ una abreviació de $r \cdot r^*$
  - r^k una abreviació de $r \cdot \ldots \cdot r$ (amb k aparicions de r)

Definició

Definició: Llenguatge associat

Anomenem llenguatge associat a una expressió regular r sobre alfabet Σ el llenguatge $L(r)$ definit com

1. $\emptyset$, $\{\lambda\}$ o $\{a\}$ si r és $\emptyset$, λ o a (amb $a \in \Sigma$), resp.
2. $L(r_1) \cup L(r_2)$ si $r = (r_1 + r_2)$
3. $L(r_1) \cdot L(r_2)$ si $r = (r_1 \cdot r_2)$
4. $(L(r_1))^*$ si $r = (r_1^*)$

Exemples

Exemples

1. $L((a^2)^*) = \{a^n \mid n \in \dot{2}\}$
2. $L((a+\lambda)b^*) = \{a\}\{b\}^* \cup \{b\}^*$
3. $L((0+\lambda)(1+\lambda)) = \{\lambda, 0, 1, 01\}$
4. $L((0+1)^*\emptyset) = \emptyset$
5. $L(\emptyset^*) = \{\lambda\}$
6. $L((01^*0+1)^*) = \{\ w \in \{0,1\}^* \mid |w|_0 \in \dot{2}\ \}$

Exemples

Exemple: $L((0+\lambda)(1+\lambda))$

Calculem el llenguatge associat aplicant la definició:

$$\begin{aligned} L((0+\lambda)(1+\lambda)) &= L(0+\lambda) \cdot L(1+\lambda) \\ &= (L(0) \cup L(\lambda)) \cdot (L(1) \cup L(\lambda)) \\ &= (\{0\} \cup \{\lambda\}) \cdot (\{1\} \cup \{\lambda\}) \\ &= \{0, \lambda\} \cdot \{1, \lambda\} \\ &= \{\lambda, 0, 1, 01\}. \end{aligned}$$

Exemples

Exemple: $(01^*0 + 1)^*$

Aplicant la definició a $r = (01^*0 + 1)^*$ obtenim:

$$\begin{aligned} L(r) &= L(01^*0 + 1)^* \\ &= (L(01^*0) \cup L(1))^* \\ &= (L(0)L(1)^*L(0) \cup L(1))^* \\ &= (\{0\}\{1\}^*\{0\} \cup \{1\})^* \end{aligned}$$

Però es pot demostrar que $L(r) = \{\ w \in \{0,1\}^* \mid |w|_0 \in \dot{2}\ \}$.

Regularitat

Teoremes

Veurem que els llenguatges associats a les expressions regulars formen exactament la classe dels regulars.

Fixem-nos que els llenguatges associats a les e.r.'s dels exemples s'han expressat en funció de $\emptyset$, $\{\lambda\}$ i $\{a\}$, per a un símbol $a \in \Sigma$ mitjançant les operacions de reunió ($+$), concatenació ($\cdot$) i estrella ($*$).

Teorema

El llenguatge associat a una e.r. és regular.

Demostració

Sigui r una e.r. sobre Σ. Llavors, $L(r)$ es pot expressar en funció dels llenguatges $\emptyset$, $\{\lambda\}$ i $\{a\}$, per a un símbol $a \in \Sigma$ a través de les operacions de reunió, concatenació i estrella.

Com que $\emptyset$, $\{\lambda\}$ i $\{a\}$ són llenguatges regulars i els regulars són tancats per les operacions esmentades, $L(r)$ ha de ser regular.

Teoremes

Teorema (Kleene, 1956)

Tot llenguatge regular és el llenguatge associat a una e.r.

Demostració (McNaughton i Yamada, 1960)

Sigui $L = L(M)$ on $M = (Q, \Sigma, \delta, q_0, F)$ és un DFA. Sigui $Q = \{q_0, \ldots, q_n\}$. Per a $0 \leq i, j \leq n$ i $1 \leq k \leq n+1$, definim

$$L_{i,j,k} = \{w \in \Sigma^* \mid q_i w = q_j \ \wedge \ \forall y, z \in \Sigma^+ \ \forall l \, (w = yz \wedge q_i y = q_l \Rightarrow l < k)\}$$

com el llenguatge dels mots que passen de q_i a q_j a través d'estats intermedis q_l t.q. $l < k$. Observem que

$$L = \bigcup_{j \mid q_j \in F} L_{0,j,n+1}.$$

Per tant, només cal veure que cada llenguatge $L_{i,j,k}$ és regular.

Teoremes

Demostració (McNaughton i Yamada, 1960)

Recordem que un mot és a $L_{i,j,k}$ si permet passar de q_i a q_j a través d'estats intermedis q_l t.q. $l < k$. Veiem que $L_{i,j,k}$ és regular per inducció en k:

- $k = 0$: El llenguatge té una e.r. associada perquè és finit:

$$L_{i,j,0} = \{a \in \Sigma \mid q_i a = q_j\} \cup \{\lambda \mid i = j\}.$$

- $k + 1$: Observem que

$$L_{i,j,k+1} = L_{i,j,k} \ \cup \ L_{i,k,k} \cdot L_{k,k,k}^* \cdot L_{k,j,k}.$$

  Per h.i., els llenguatges de la dreta tenen les e.r.'s associades resp.

$$r_{i,j,k}, r_{i,k,k}, r_{k,k,k}, \text{ i } r_{k,j,k}.$$

  Per tant, $L_{i,j,k+1} = L(r_{i,j,k} + r_{i,k,k} \cdot r_{k,k,k}^* \cdot r_{k,j,k})$.

Lema d'Arden

El mètode de la demostració de McNaughton i Yamada és difícil d'aplicar a mà. Una alternativa pràctica és la basada en el lema d'Arden.

Definició

Una equació lineal (per la dreta) sobre un alfabet Σ és una expressió de la forma

$$X = AX + B$$

on A i B són llenguatges sobre Σ.

Un llenguatge L és solució de l'equació si $L = A \cdot L \cup B$.

Notació

Descriurem sovint els llenguatges de les equacions lineals mitjançant e.r.'s sense aplicar l'operador de *llenguatge associat*.

Per exemple, $X = (01^*0 + 1)X + \lambda$ és una equació lineal.

Lema d'Arden

Exemple

Considerem l'equació lineal:

$$X = (01^*0 + 1)X + \lambda.$$

Veiem que, si L és solució de l'equació, aleshores

1. $\lambda \in L$
2. Pel punt anterior, $01^*0 + 1 = (01^*0 + 1) \cdot \lambda \subseteq L$.
3. Per a tot $k \geq 0$, $(01^*0 + 1)^k \subseteq L$ per substitució repetida en l'equació.
4. Per tant, $(01^*0 + 1)^* \subseteq L$.

Ara generalitzem el raonament de l'exemple anterior.

Lema d'Arden

Lema d'Arden

Considerem l'equació $X = AX + B$. Aleshores,

1. El llenguatge A^*B n'és solució.
2. Qualsevol altra solució conté A^*B.
3. Si $\lambda \notin A$, aleshores A^*B n'és l'única solució.

Lema d'Arden

Lema d'Arden

Considerem l'equació $X = AX + B$. Aleshores,

1. El llenguatge A^*B n'és solució.

Demostració

Substituint X per A^*B tenim

$$A(A^*B) + B = A^+B + B = (A^+ + \lambda)B = A^*B.$$

Lema d'Arden

Lema d'Arden

Considerem l'equació $X = AX + B$. Aleshores,

2. Qualsevol altra solució conté A^*B.

Demostració

Sigui L una solució qualsevol de l'equació. Per a tot $k \in \mathbb{N}$ tenim

$$\begin{aligned} L &= AL + B \\ &= A^2L + AB + B \\ &= \dots \\ &= A^{k+1}L + A^kB + \dots AB + B \end{aligned}$$

i, per tant, $A^kB \subseteq L$, d'on es dedueix que $A^*B \subseteq L$.

Lema d'Arden

Lema d'Arden

Considerem l'equació $X = AX + B$. Aleshores,

3. Si $\lambda \notin A$, aleshores A^*B n'és l'única solució.

Demostració

Suposem que l'equació té L i M com a solucions i, per tant,

$$L = AL + B \quad \text{i} \quad M = AM + B.$$

Veurem que tot mot w de L és a M per inducció en $n = |w|$:

- $n = 0$. Llavors $w = \lambda$. Si $w \in L$, tenim $w \in B$ i, per tant, $w \in M$.
- $n > 0$. Si $w \in L$, llavors $w \in AL$ o $w \in B$. En el segon cas, $w \in M$. En el primer, $w = xv$ amb $x \in A$ i $v \in L$. Com que $\lambda \notin A$, tenim que $|v| < n$ i, per h.i., $v \in M$. Així, $w \in AM \subseteq M$.

Per tant, $L \subseteq M$. Simètricament, $M \subseteq L$ i, per tant, $L = M$.

Lema d'Arden

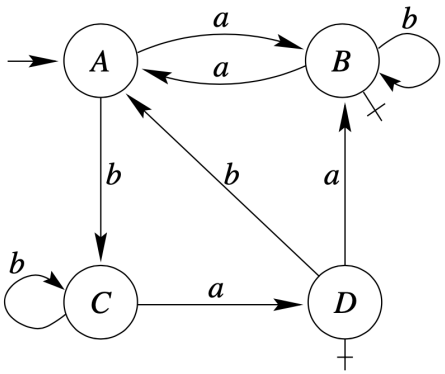
Una alternativa per construir una e.r. a partir d'un DFA és:

1. Plantejar un sistema d'equacions lineals on cada variable descriu el llenguatge dels mots generats a partir d'un estat de l'autòmat
2. Resoldre el sistema d'equacions amb l'ajut del lema d'Arden
3. Obtenir la solució de la variable associada a l'estat inicial

Lema d'Arden

Exemple (secció 6.4, Cases-Màrquez)

Considereu el DFA, M, de la figura següent.



$$
\begin{aligned}
(1) \quad & A = aB + bC \\
(2) \quad & B = aA + bB + \Lambda \\
(3) \quad & C = aD + bC \\
(4) \quad & D = aB + bA + \Lambda\,.
\end{aligned}
$$

Lema d'Arden

Exemple (secció 6.4, Cases-Màrquez)

$$
\begin{aligned}
&(1) \quad A = aB + bC \\
&(2) \quad B = aA + bB + \Lambda \\
&(3) \quad C = aD + bC \\
&(4) \quad D = aB + bA + \Lambda\,.
\end{aligned}
$$

Resolent per Arden les equacions 2 i 3 obtenim

$$B = b^*(aA + \Lambda) = b^*aA + b^*$$

i

$$C = b^*aD,$$

respectivament. Substituint la nova expressió de B a l'equació 4 obtenim

$$D = ab^*(aA + \Lambda) + bA + \Lambda = (ab^*a + b)A + (ab^* + \Lambda)\,.$$

Substituint aquesta expressió de D a la de C anterior obtenim

$$C = (b^*aab^*a + b^*ab)A + (b^*aab^* + b^*a)\,,$$

i substituint B i C a l'equació 1 obtenim l'equació lineal següent:

$$
\begin{aligned}
A &= ab^*aA + ab^* + (bb^*aab^*a + bb^*ab)A + (bb^*aab^* + bb^*a) \\
&= (ab^*a + bb^*aab^*a + bb^*ab)A + (ab^* + bb^*aab^* + bb^*a)\,.
\end{aligned}
$$

Finalment, per Arden, obtenim l'expressió regular buscada

$$\mathsf{L}(M) = A = (ab^*a + bb^*aab^*a + bb^*ab)^*(ab^* + bb^*aab^* + bb^*a)\,.$$

Regularitat

Limitacions dels DFAs

Veurem com demostrar que alguns llenguatges no poden ser reconeguts per cap DFA. Considerem el llenguatge

$$\{a^n b^n \mid n \geq 0\}.$$

Un DFA que el vulgui reconèixer hauria de comptar el nombre d'as que han aparegut abans de les bes per poder decidir correctament. Però el nombre d'as és arbitràriament gran!

En aquest apartat veurem una eina per demostrar que un llenguatge no és regular.

Limitacions dels DFAs

Ens podem preguntar si la necessitat de comptar un nombre arbitrari de símbols no és prova suficient de no regularitat. Els exemples següents mostren que la resposta és **no** i que necessitem una eina formal:

1. $A = \{w \in \{a, b\}^* \mid |w|_a = |w|_b\}$
2. $B = \{w \in \{a, b\}^* \mid |w|_{ab} = |w|_{ba}\}$

El llenguatge A no és regular però sorprenentment B sí és regular (per què?).

Lema de bombament (Bar-Hillel *et al*, 1961, veure Cases-Màrquez 7.1)

Si L és un llenguatge regular, existeix un natural $N \geq 1$ tal que tot mot $w \in L$ amb $|w| \geq N$ admet una factorització $w = xyz$ que satisfà

1. $|xy| \leq N$,
2. $|y| \geq 1$ i
3. $xy^i z \in L$ per a tot $i \geq 0$.

Condició de bombament

Definim la condició de bombament per a un llenguatge L:

$$\mathcal{B}(L) \equiv \exists N \geq 1 \;\; \forall w \in L \;\; |w| \geq N \;\wedge\; \exists x, y, z$$

$$(w = xyz \wedge |xy| \leq N \;\wedge |y| \geq 1) \Rightarrow \forall i \geq 0 \;\; xy^i z \in L.$$

Lema de bombament

$L \in \mathrm{REG} \Rightarrow \mathcal{B}(L)$.

Lema de bombament (Bar-Hillel *et al*, 1961, veure Cases-Màrquez 7.1)

Si L és un llenguatge regular, existeix un natural $N \geq 1$ tal que tot mot $w \in L$ amb $|w| \geq N$ admet una factorització $w = xyz$ que satisfà

1. $|xy| \leq N$,
2. $|y| \geq 1$ i
3. $xy^i z \in L$ per a tot $i \geq 0$.

Condició de bombament

Definim la condició de bombament per a un llenguatge L:

$$\mathcal{B}(L) \equiv \exists N \geq 1 \;\; \forall w \in L \;\; |w| \geq N \;\wedge\; \exists x, y, z$$

$$(w = xyz \wedge |xy| \leq N \;\wedge |y| \geq 1) \Rightarrow \forall i \geq 0 \;\; xy^i z \in L.$$

Lema de bombament

$L \in \mathrm{REG} \Rightarrow \mathcal{B}(L)$.

La manera com aplicarem el lema és a través del contrarecíproc, equivalent al lema, que dona una condició de no regularitat.

Lema de bombament

$L \in \text{REG} \Rightarrow \mathcal{B}(L)$.

Contrarecíproc del lema de bombament

$\neg\mathcal{B}(L) \Rightarrow L \notin \text{REG}$.

Negació de la condició de bombament

Aquesta és, per tant, la condició que cal comprovar per deduir que un llenguatge L no és regular:

$$\neg\mathcal{B}(L) \equiv \forall N \geq 1 \;\; \exists w \in L \;\; |w| \geq N \;\wedge\; \forall x, y, z$$

$$(w = xyz \wedge |xy| \leq N \;\wedge |y| \geq 1) \Rightarrow \exists i \geq 0 \;\; xy^i z \notin L.$$

Exemple (Cases-Màrquez, exemple 7.2)

Volem veure que el llenguatge dels palíndroms parells no és regular:

$$L = \{ww^R \mid w \in \{a,b\}^*\}.$$

1. Donat un $N \geq 1$, triem el mot $w = a^N bba^N \in L$, on $|w| \geq N$.
2. Donada una factorització $w = xyz$ amb $|xy| \leq N$ i $|y| \geq 1$, per la forma de w sabem que existeixen j, k amb $j + k \leq N$ i $k \geq 1$ tals que

$$x = a^j, \quad y = a^k, \quad z = a^{N-j-k} bba^N.$$

3. Triant un bombament $i \neq 1$, es pot deduir que $xy^i z \notin L$. Per exemple, per a $i = 2$ tenim

$$xy^2 z = a^j a^{2k} a^{N-j-k} bba^N = a^{N+k} bba^N.$$

Com que $k \geq 1$, el submot bb no ocupa la posició central de $xy^2 z$ i, per tant, $xy^2 z$ no és palíndrom, és a dir, $xy^2 z \notin L$.

El recíproc del lema és fals.

Recíproc del lema de bombament

$\mathcal{B}(L) \Rightarrow L \in \text{REG}$.

Demostració (veure Cases-Màrquez. secció 7.1)

Sigui $L = \{uu^R v \mid u, v \in \{a, b\}^+\}$. Es divideix la demostració en dues parts:

- Es demostra que L compleix la condició de bombament, és a dir, que $\mathcal{B}(L)$ és cert.
- Es demostra que L no és regular aplicant el contrarecíproc del lema de bombament a

$$L' = L \cap L(aba^* bba^* bab) = \{aba^n bba^n bab \mid n \geq 0\}.$$

Com que la intersecció de regulars és regular i el segon llenguatge intersecat ho és, es dedueix que el primer, L, no pot ser-ho.