I metodi formali dell'Analisi Lessicale: Le Espressioni Regolari (ER)

N.Fanizzi - V.Carofiglio

6 aprile 2016

Espressioni Regolari

Dato un alfabeto finito X, una espressione regolare è una notazione per decrivere linguaggi regolari sull'alfabeto ampliato $X \cup \{\lambda, +, *, \cdot, \emptyset, (,)\}$:

ESEMPI

- $L = \{a\} \quad R(L) = a$
- $L = \{a, b, c\} \quad R(L) = a + b + c$

  (+ denota UNIONE tra linguaggi)
- $L = (\{a\} \bigcup \{bc\})^* \quad R(L) = (a + b \cdot c)$

  (* e $\cdot$ denotano ITERAZIONE e CONCATENAZIONE tra linguaggi)

Espressioni Regolari - definizione

Dato un alfabeto finito X, una stringa R sull'alfabeto ampliato $X \cup \{\lambda, +, *, \cdot, \emptyset, (,)\}$ è una **espressione regolare** di alfabeto X sse vale una delle seguenti condizioni:

- $R = \emptyset$
- $R = \lambda$
- $R = a$ con $a \in X$
- $R = (R_1 + R_2)$ con R_1, R_2 espressioni regolari di alfabeto X
- $R = (R_1 \cdot R_2)$ con R_1, R_2 espressioni regolari di alfabeto X
- $R = (R_1)^*$ con R_1 espressione regolare di alfabeto X

L'**insieme delle espressioni regolari** di alfabeto X viene denotato con $\mathcal{R}$

Linguaggi Regolari

Dato un alfabeto finito X
un linguaggio L su X è un **linguaggio regolare** sse:

- L è finito
  oppure
- L può essere ottenuto induttivamente mediante le operazioni:
  1. $L = L_1 \cup L_2$ con L_1, L_2 regolari
  2. $L = L_1 \cdot L_2$ con L_1, L_2 regolari
  3. $L = L_1^*$ con L_1 regolare

Osservazione: $\emptyset$ e $\{\lambda\}$ sono linguaggi regolari
Definiamo l'insieme di tali linguaggi come **classe dei linguaggi regolari** denotata con $\mathcal{L}_{REG}$

Operazioni sui linguaggi (Dovere di cronaca)

Dati due linguaggi L e L' definiti sullo stesso alfabeto X:

unione $L \cup L' = \{w \in X^* \mid w \in L \vee w \in L'\}$

prodotto $L \cdot L' = \{w = w_1 \cdot w_2 \in X^* \mid w_1 \in L \wedge w_2 \in L'\}$

iterazione $L^* = \{w_1 \dots w_n \in X^* \mid \forall n \forall i : \ w_i \in L, \ 0 \leq i \leq n\}$

complemento $\overline{L} = \{w \in X^* \mid w \notin L\}$

intersezione $L \cap L' = \{w \in X^* \mid w \in L \wedge w \in L'\}$

potenza $L^k = \begin{cases} \{\lambda\} & \text{k=0} \\ L^{k-1} \cdot L & \text{k>0} \end{cases}$

chiusura (transitiva) $L^+ = \bigcup_{k>0} L^k$
(quindi si può scrivere $L^* = L^0 \cup L^+ = \{\lambda\} \cup L^+$)

Esempi.

Dati i linguaggi $L_1 = \{a^{2n} \mid n \geq 0\}$ e $L_2 = \{b, cc\}$

- $L_1 \cdot L_2 = \{b, cc, aab, aacc, aaaab, aaaacc, \ldots\}$
- $L_2 \cdot L_1 = \{b, cc, baa, ccaa, baaaa, ccaaaa, \ldots\}$
- $L_1 \cup L_2 = L_2 \cup L_1 = \{\lambda, b, cc, aa, aaaa, aaaaaa, \ldots\}$
- $L_1^* = \{\lambda, aa, aaaa, aaaaaa, aaaaaaaa, \ldots\}$
- $L_2^* = \{\lambda, b, cc, bb, bcc, ccb, bbb, cccc, \ldots\}$
- $L_2^0 = \{\lambda\}$
  $L_2^1 = \{b, cc\}$
  $L_2^2 = \{bb, bcc, ccb, cccc\} \ldots$
- $L_2^+ = \{b, cc, bb, bcc, ccb, bbb, cccc, \ldots\}$

Corrispondenza Linguaggi - Espressioni Regolari

Ad ogni espressione regolare R
corrisponde un linguaggio regolare $S(R)$:

espressione regolare	linguaggio regolare
$\emptyset$	$\emptyset$
λ	$\{\lambda\}$
a	$\{a\}$
$(R_1 + R_2)$	$S(R_1) \cup S(R_2)$
$(R_1 \cdot R_2)$	$S(R_1) \cdot S(R_2)$
$(R)^*$	$(S(R))^*$

Proposizione.
Un linguaggio su X è regolare sse esso corrisponde ad una espressione regolare di alfabeto X
ossia definita la funzione $S : \mathcal{R} \longrightarrow \wp(X^*)$, si ha che:

$$\mathcal{L}_{REG} = \{L \in \wp(X^*) \mid \exists R \in \mathcal{R}: \ \ L = S(R)\}$$

Osservazioni.

- un linguaggio regolare può essere descritto da più di una espressione regolare, cioè
- la funzione S non è inettiva

Due **espressioni regolari** R_1 e R_2 si dicono **equivalenti**, e si scrive $R_1 = R_2$, sse

$$S(R_1) = S(R_2)$$

Esempio.
L linguaggio delle parole su $X = \{a, b\}$ con a e b alternate, inizianti e terminanti con b
L descrivibile attraverso l'espressione

$$R_1 = b(ab)^*$$

ovvero con

$$R_2 = (ba)^* b$$

pertanto $R_1 = R_2$

1.	$(R_1 + R_2) + R_3 = R_1 + (R_2 + R_3)$	prop. associativa di $+$
2.	$R_1 + R_2 = R_2 + R_1$	prop. commutativa di $+$
3.	$R + \emptyset = \emptyset + R = R$	$\emptyset$ elem. neutro per $+$
4.	$R + R = R$	Idempotenza
5.	$(R_1 \cdot R_2) \cdot R_3 = R_1 \cdot (R_2 \cdot R_3)$	prop. associativa di $\cdot$
6.	$R_1 \cdot R_2 \neq R_2 \cdot R_1$	non commutatività di $\cdot$
7.	$R \cdot \lambda = \lambda \cdot R = R$	λ elem. neutro per $\cdot$
8.	$R \cdot \emptyset = \emptyset \cdot R = \emptyset$	$\emptyset$ elem. assorbente per $\cdot$
9.	$R_1 \cdot (R_2 + R_3) = (R_1 \cdot R_2) + (R_1 \cdot R_3)$	prop. distributiva di $\cdot$ risp. a $+$
10.	$(R_1 + R_2) \cdot R_3 = (R_1 \cdot R_3) + (R_2 \cdot R_3)$	prop. distributiva di $\cdot$ risp. a $+$
11.	$(R)^* = (R)^* \cdot (R)^* = ((R)^*)^* = (\lambda + R)^*$	
12.	$(\emptyset)^* = (\lambda)^* = \lambda$	
13.	$(R)^* = \lambda + R + \ldots + R^n + (R^{n+1} \cdot R^*)$	
14.	$(R_1 + R_2)^* = (R_1^* + R_2^*)^* = (R_1^* \cdot R_2^*)^* =$	
	$= (R_1^* \cdot R_2^*)^* \cdot R_1^* = R_1^* \cdot (R_1^* \cdot R_2^*)^*$	
15.	$(R_1 + R_2)^* \neq R_1^* + R_2^*$	in generale
16.	$R \cdot R^* = R^* \cdot R$	
17.	$R_1 \cdot (R_2 \cdot R_1)^* = (R_1 \cdot R_2)^* \cdot R_1$	
18.	$(R_1^* \cdot R_2)^* = \lambda + (R_1 + R_2)^* \cdot R_2$	
19.	$(R_1 \cdot R_2^*)^* = \lambda + R_1 \cdot (R_1 + R_2)^*$	
20.	$R_1 = R_2 \cdot R_1 + R_3$ sse $R_1 = R_2^* \cdot R_3$	con $R_2 \neq \lambda$
	$R_1 = R_1 \cdot R_2 + R_3$ sse $R_1 = R_3 \cdot R_2^*$	

Esercizi

1. Data l'espressione $R = a^* \cdot (a + b)$, trovare il linguaggio associato.
2. Data L'espressione $R = (aa)^*(bb)^*b$, trovare il linguaggio associato
3. Data l'espressione $R = (a + b)^*(a + bb)$, trovare il linguaggio associato
4. Fissato l'alfabeto $X = \{0, 1\}$ dare l'espressione regolare R, tale che:
   - $S(R) = \{w \in X^* | \ w \ ha \ almeno \ una \ coppia \ di \ 0 \ consecutivi\}$
   - $S(R) = \{w \in X^* | \ w \ non \ ha \ coppie \ di \ 0 \ consecutivi\}$
5. Trovare tutte le stringhe di lunghezza minore di quattro in $S((a + b)^*b(a + ab)^*)$
6. Trovare l'espressione regolare per:
   - $L = \{a^n b^m \mid (n + m) pari\}$
   - $L = \{a^n b^m \mid n \geq 4; \ m \geq 3\}$

Esercizi

1. Scrivere le espressioni regolari per i seguenti linguaggi su $X = \{0, 1\}$
   - tutte le stringhe che terminano per 01
   - tutte le stringhe che non terminano per 01
   - tutte le stringhe contenenti un numero pari di 0
   - tutte le stringhe che hanno almeno due occorrenze di sottostringhe 00 (nota che 000 contiene due occorrenze di 00)
   - tutte le stringhe che hanno al piú due occorrenze di sottostringhe 00
   - tutte le stringhe che non contengono occorrenze di 101

Esercizi

1. Trovare le espressioni regolari per i seguenti linguaggi su $\{a, b\}$
   - $L = \{w \mid |w| \; mod \; 3 \; = \; 0\}$
   - $L = \{w \mid n_a(w) \; mod \; 3 \; = \; 0\}$
   - $L = \{w \mid n_a(w) \; mod \; 5 \; < \; 0\}$

Esercizio1

Individuare il linguaggio $S(a^* \cdot (a+b))$

$$
\begin{gathered}
S(a^* \cdot (a+b)) = \\
S(a^*) \cdot S(a+b) = \\
(S(a))^* \cdot (S(a) \cup S(b)) = \\
\{\lambda, a, aa, aaa, \ldots\} \cdot \{a, b\} \\
\{\lambda, a, aa, aaa, \ldots, b, ab, aab, \ldots\}
\end{gathered}
$$

Esercizio2

Individuare il linguaggio $S((a+b)^*(a+bb))$

$(a+b)^*$ stringhe di "a" e "b"

$(a+bb)$ stringa "a" oppure stringa "bb"

$S((a+b)^*(a+bb))$ é l'insieme su $\{a,b\}$ di stringhe che terminano per "a" oppure per "bb".

ovvero:

$$S((a+b)^*(a+bb)) = \{a, bb, aa, abb, ba, bbb, ...\}$$

Esercizio3

Individuare il linguaggio $S((aa)^*(bb)^*b)$

$$S((aa)^*(bb)^*b) = \{a^{2n}b^{2m+1} \mid n \geq 0,\ m \geq 0\}$$

Esercizio4a

Individuare l'espressione regolare per il linguaggio su $X\{0,1\}$

$$L = \{w \in X^* \mid w \textit{ ha almeno una coppia di zeri consecutivi}\}$$

Ogni stringa di L deve contenere:

- almeno un "00"
- ed é del tipo "x00y", con x e y arbitrarie stringhe su X^* $((0+1)^*)$

dunque $L = S((0+1)^*\ 00\ (0+1)^*)$

Esercizio4b

Individuare l'espressione regolare per il linguaggio

$$L = \{w \in \{0,1\}^* \mid w \; NON \; ha \; coppie \; di \; zeri \; consecutivi\}$$

Osserviamo che

- se occorre uno "0", subito dopo ci deve essere un "1"
- la sotto-stringa "01" puó essere preceduta e seguita da"1"

Dunque:
La risposta contiene ripetiz. di "1....101.....1" (ovvero $(1^*(01)1^*)^*$

- mancano le stringhe del tipo "1...101...10" $((1^*(01)1^*0)$

Dunque:
La risposta contiene anche ripetizioni di stringhe della forma "1....101.....10" (ovvero $(1^*(01)1^*)^* \; (\lambda + 0)$

- mancano le stringhe del tipo "1...10" $((1^*0))$
- mancano le stringhe del tipo "1...1" $((1^*))$

dunque mancano $1^*(0+\lambda)$

dunque
$L = S(\; (1^*(01)1^*)^* \; (0+\lambda) \; + \; (1^*(0+\lambda)) \;)$

Esercizio4b-bis

Individuare l'espressione regolare per il linguaggio

$$L = \{w \in \{0,1\}^* \mid w \ NON \ ha \ coppie \ di \ zeri \ consecutivi\}$$

- $S(\ (1^*(01)1^*)^*\ (0+\lambda)\ +\ (1^*(0+\lambda))\ )$
- $S(\ (1+(01))^*\ (0+\lambda)\ )$

definiscono lo stesso linguaggio?