Expresiones regulares y derivadas

Teoría de Lenguajes

1^{er} cuatrimestre de 2002

1 Expresiones regulares

Las expresiones regulares son expresiones que se utilizan para denotar lenguajes regulares. No sirven para denotar lenguajes menos restrictivos que los regulares (libres de contexto y demás). Se definen recursivamente de la siguiente manera:

1. La expresión regular $\emptyset$ define el lenguaje vacío.

   La expresión regular λ define el lenguaje $\{\lambda\}$.

   La expresión regular a define el lenguaje {a}.

2. Sean:

   v la expresión regular que define el lenguaje L_v

   w la expresión regular que define el lenguaje L_w

   entonces:

   (a) la expresión regular $(v.w)$ define el lenguaje $L_v.L_w$

   (b) la expresión regular $(v|w)$ define el lenguaje $L_v \cup L_w$

   (c) la expresión regular v^* define el lenguaje L_v^*

   (d) la expresión regular v^+ define el lenguaje L_v^+

Ejemplos:

- La expresión regular $(a|b)^+$ define el lenguaje de todas las cadenas de as y bs de longitud mayor o igual que 1.
- La expresión regular $(1|\ldots|9)(0|1|\ldots|9)^*|0$ define todos los números naturales sin ceros no significativos.

2 Derivada de un lenguaje

Dado un lenguaje L sobre un alfabeto Σ definimos las siguientes derivadas:

1. $\partial\lambda(L) = L$
2. $\partial \mathrm{a}(L) = \{\alpha : \mathrm{a}\alpha \in L\}$ donde $\mathrm{a} \in \Sigma, \alpha \in \Sigma^*$
3. $\partial w(L) = \partial u(\partial \mathrm{a}(L))$ si $w = \mathrm{a}.u, \mathrm{a} \in \Sigma, u \in \Sigma^*$

Notaremos $\partial \mathrm{ab}\ L = \partial \mathrm{b}(\partial \mathrm{a}(L))$

Por ejemplo, si $L = \{\mathrm{a}^{3n} : n \in \mathbf{N}_{\geq 1}\} = \{\mathrm{aaa}, \mathrm{aaaaaa}, \mathrm{aaaaaaaaa}, \ldots\}$, entonces:

$$\begin{aligned} \partial \mathrm{a}\ L &= \{\mathrm{a}^{3n-1} : n \in \mathbf{N}_{\geq 1}\} = \{\mathrm{aa}, \mathrm{aaaaa}, \mathrm{aaaaaaaa}, \ldots\} \\ \partial \mathrm{b}\ L &= \emptyset \\ \partial \mathrm{aaa}\ L &= L \cup \{\lambda\} \\ \partial \mathrm{aaaa}\ L &= \partial \mathrm{a} L \end{aligned}$$

Para un lenguaje L definimos también $\epsilon(L)$, que indica si la cadena nula pertenece o no al lenguaje.

$$\epsilon(L) = \begin{cases} \emptyset & \text{si} \quad \lambda \notin L \\ \{\lambda\} & \text{si} \quad \lambda \in L \end{cases}$$

3 Derivada de una expresion regular

De la misma manera que definimos la derivada de cadenas de un lenguaje, podemos definir la derivada $\partial x(u)$ de una expresión regular u con respecto a una cadena $x \in \Sigma^*$, mediante las siguientes reglas:

1. $\partial\lambda(u) = u$
2. Para $\mathrm{a} \in \Sigma$:

$$\begin{aligned} \partial \mathrm{a}(\emptyset) &= \emptyset \\ \partial \mathrm{a}(\lambda) &= \emptyset \\ \partial \mathrm{a}(\mathrm{b}) &= \begin{cases} \lambda & \text{si} \quad \mathrm{a} = \mathrm{b} \\ \emptyset & \text{si} \quad \mathrm{a} \neq \mathrm{b} \end{cases} \end{aligned}$$

3. Sean u y v expresiones regulares que denotan los lenguajes L_u y L_v respectivamente, entonces:

$$\begin{aligned} \partial \mathrm{a}(u|v) &= \partial \mathrm{a}(u)|\partial \mathrm{a}(v) \\ \partial \mathrm{a}(u.v) &= \partial \mathrm{a}(u).v|\epsilon(u).\partial \mathrm{a}(v) \\ \partial \mathrm{a}(u^*) &= \partial \mathrm{a}(u).u^* \end{aligned}$$

donde:

$$\epsilon(u) = \begin{cases} \emptyset & \text{si} \quad \lambda \notin L_u \\ \lambda & \text{si} \quad \lambda \in L_u \end{cases}$$

4. Para $\mathrm{a} \in \Sigma \wedge x \in \Sigma^*$:

$$\partial \mathrm{a}x\ u = \partial x(\partial \mathrm{a}\ u)$$

Ejemplos:

$\partial \mathrm{a}(\mathrm{a}) = \lambda$
$\partial \mathrm{a}(\mathrm{a}|\mathrm{b}) = \partial \mathrm{a}(\mathrm{a})|\partial \mathrm{a}(\mathrm{b}) = \lambda|\emptyset = \lambda$
$\partial \mathrm{a}((\mathrm{a}|\mathrm{b}).\mathrm{c}) = \partial \mathrm{a}(\mathrm{a}|\mathrm{b}).\mathrm{c}|\epsilon(\mathrm{a}|\mathrm{b}).\partial \mathrm{a}(\mathrm{c}) = \lambda.\mathrm{c}|\emptyset.\emptyset = c$
$\partial \mathrm{a}(\mathrm{a}^*) = \mathrm{a}^*$
$\partial \mathrm{a}(\mathrm{a}^+) = \partial \mathrm{a}(\mathrm{a}.\mathrm{a}^*) = \partial \mathrm{a}(\mathrm{a}).\mathrm{a}^*|\epsilon(\mathrm{a}).\partial \mathrm{a}(\mathrm{a}^*) = \lambda.\mathrm{a}^*|\emptyset.\mathrm{a}^* = \mathrm{a}^*$

Dada una expresión regular, ésta denota un lenguaje regular, para el cual existe un autómata finito que lo reconoce.

Sea $A = (Q = \{q_0, q_1, \ldots, q_n\}, \Sigma = \{a_1, a_2, \ldots, a_m\}, F \subseteq Q, q_0, \delta_A)$ un autómata finito determinístico.

δ_A es la función de transición y está definida en:

$$\delta_A : Q \times \Sigma \to Q$$

Esta función se puede extender para cadenas de cualquier longitud sobre Σ, definiendo la función δ:

$$\delta : Q \times \Sigma^* \to Q \text{ tal que } \begin{cases} \delta(q, \mathrm{a}x) & = & \delta(\delta_A(q, \mathrm{a}), x), \text{ para } x \in \Sigma^* \\ \delta(q, \lambda) & = & q \end{cases}$$

El lenguaje aceptado por el autómata se define como:

$$L(A) = \{x \in \Sigma^* \ | \delta(q_0, x) \in F\}$$

De la misma forma que definimos el lenguaje L como el conjunto de cadenas aceptadas por el autómata partiendo del estado q_0, podemos definir un lenguaje L_i para un estado q_i dado del autómata, que corresponde al lenguaje aceptado por el autómata si partiéramos, no de q_0 sino de ese estado q_i:

$$L_i = \{x \in \Sigma^* \ | \delta(q_i, x) \in F\}$$

Veamos el siguiente ejemplo. Dado el autómata $A = (\{q_0, q_1, q_2, q_3\}, \{\mathrm{a}, \mathrm{b}\}, \{q_1, q_3\}, q_0, \delta_1)$

δ_1	a	b
0	1	2
1	1	3
2	3	
3	3	

	a	b
0	1	2
1	1	3
2	3	T
3	3	T
T	T	T

Del lado derecho figura el autómata completado con el estado trampa, que es un estado que se puede agregar al autómata para completar las transiciones no definidas. Trampa es un estado no final, adonde van a parar las transiciones no definidas y que cicla sobre el mismo con cada símbolo de Σ^*.

Partiendo del estado q_0, tenemos el lenguaje $L = L_0$ aceptado por el autómata, que está definido por la expresión regular $(\mathrm{a}^+(\mathrm{ba}^*|\lambda)|\mathrm{ba}^+)$. Ahora, si en vez de partir del estado q_0 partiéramos del estado q_1, tendríamos el lenguaje L_1, que corresponde a la expresión regular $\mathrm{a}^*\mathrm{ba}^*$, y si partimos del estado q_2 tenemos el lenguaje L_2 con expresión regular a^+. Análogamente, $L_3 = \mathrm{a}^*$ y el estado trampa define el lenguaje vacío, ya que partiendo de él no se puede reconocer ninguna cadena, $L_T = \emptyset$.

Volviendo al autómata general definido más arriba, veamos que para un estado q_i dado $(1 \leq i \leq n)$, y para cada símbolo $a_j \in \Sigma$ $(1 \leq j \leq m)$, podemos armar el siguiente sistema de ecuaciones:

$$\begin{array}{rcll} \delta(q_i, a_1) & = & p_{k_1} & , p_{k_1} \in Q \\ \delta(q_i, a_2) & = & p_{k_2} & , p_{k_2} \in Q \\ \ldots & & & \\ \delta(q_i, a_m) & = & p_{k_m} & , p_{k_m} \in Q \end{array} \tag{1}$$

Basándonos en estas ecuaciones, podemos definir el lenguaje L_i de la siguiente forma:

$$L_i = a_1.L_{k_1}|a_2.L_{k_2}|\ldots|a_m.L_{k_m}|\epsilon(L_i) \tag{2}$$

En el autómata del ejemplo, estas ecuaciones corresponderían a:

$$\begin{array}{rcl} L_0 & = & \mathrm{a}.L_1|\mathrm{b}.L_2 \\ L_1 & = & \mathrm{a}.L_1|\mathrm{b}.L_3|\lambda \\ L_2 & = & \mathrm{a}.L_3|\mathrm{b}.L_T \\ L_3 & = & \mathrm{a}.L_3|\lambda \\ L_T & = & \mathrm{a}.L_T|\mathrm{b}.L_T|\emptyset \end{array}$$

Podemos resolver este sistema de ecuaciones utilizando la siguiente propiedad de los lenguajes definidos sobre Σ: (dados R, S, T lenguajes sobre Σ)

$$\text{Si } R = S.R|T \wedge \lambda \notin S, \text{ entonces } R = S^*.T$$

Utilizando esta propiedad y otras de las expresiones regulares, despejemos las ecuaciones planteadas arriba:

$L_T = (\mathrm{a}|\mathrm{b}).L_T|\emptyset$, y como $\lambda \notin \{\mathrm{a}, \mathrm{b}\}$, entonces $L_T = (\mathrm{a}|\mathrm{b})^*.\emptyset = \emptyset$
$L_3 = \mathrm{a}.L_3|\lambda$, entonces $L_3 = \mathrm{a}^*.\lambda = \mathrm{a}^*$
$L_2 = \mathrm{a}.L_3|\mathrm{b}.L_T = \mathrm{a}.\mathrm{a}^*|\emptyset = \mathrm{a}^+$
$L_1 = \mathrm{a}.L_1|\mathrm{b}.L_3|\lambda = \mathrm{a}.L_1|(\mathrm{b}.\mathrm{a}^*|\lambda)$, y como $\lambda \notin \{\mathrm{a}\}$, entonces $L_1 = \mathrm{a}^*.(\mathrm{b}.\mathrm{a}^*|\lambda)$
$L_0 = \mathrm{a}.L_1|\mathrm{b}.L_2 = \mathrm{a}^+(\mathrm{b}.\mathrm{a}^*|\lambda)|\mathrm{ba}^+$

De esta manera, tenemos un método para, dado un autómata finito, hallar la expresión regular que define el lenguaje aceptado por el autómata. Veamos que utilizando estos mismos conceptos, podemos formular un método para, dada la expresión regular que define un lenguaje, encontrar el autómata finito que acepta ese lenguaje.

Primero observemos que las ecuaciones (1) se pueden formular en términos de las derivadas del lenguaje L_i:

$$\begin{aligned} \partial a_1(L_i) &= L_{k_1} \\ \partial a_2(L_i) &= L_{k_2} \\ &\dots \\ \partial a_m(L_i) &= L_{k_m} \end{aligned}$$

Reemplazando estos valores en (2) nos queda:

$$L_i = a_1.\partial a_1(L_i)|a_2.\partial a_2(L_i)|\dots|a_m.\partial a_m(L_i)|\epsilon(L_i)$$

Si tenemos la expresión regular que define L, tenemos la expresión regular que define L_0. Si derivamos esta expresión regular con respecto a un símbolo $a_j \in \Sigma$, obtendremos la expresión regular que define el lenguaje aceptado por el autómata comenzando por el estado $\delta(q_0, a_j)$ (puede ser el mismo L_0 u otro L_δ)

Veamos cómo obtenemos el autómata que corresponde a la expresión regular $\mathrm{a}^+(\mathrm{b.a}^*|\lambda)|\mathrm{ba}^+ = L_0$:

$$\begin{aligned} \partial\mathrm{a}\ L_0 &= \partial\mathrm{a}(\mathrm{a}^+(\mathrm{b.a}^*|\lambda))|\partial\mathrm{a}(\mathrm{ba}^+) = \\ &\quad \mathrm{a}^*(\mathrm{b.a}^*|\lambda)|\emptyset.\partial\mathrm{a}(\mathrm{ba}^*|\lambda)|\emptyset = \mathrm{a}^*(\mathrm{b.a}^*|\lambda) = L_1 \\ \partial\mathrm{b}\ L_0 &= \partial\mathrm{b}(\mathrm{a}^+(\mathrm{b.a}^*|\lambda)|\mathrm{ba}^+) = \\ &\quad \partial\mathrm{b}(\mathrm{a}^+).(\mathrm{b.a}^*|\lambda)|\emptyset.\partial\mathrm{b}(\mathrm{b.a}^*|\lambda)|\partial\mathrm{b}(\mathrm{ba}^+) = \mathrm{a}^+ = L_2 \\ \partial\mathrm{a}\ L_1 &= \partial\mathrm{a}(\mathrm{a}^*(\mathrm{b.a}^*|\lambda)) = \partial\mathrm{a}(\mathrm{a}^*).(\mathrm{b.a}^*|\lambda)|\lambda.\partial\mathrm{a}(\mathrm{b.a}^*|\lambda) = \\ &\quad \mathrm{a}^*(\mathrm{b.a}^*|\lambda)|\lambda.\emptyset = \mathrm{a}^*(\mathrm{b.a}^*|\lambda) = L_1 \\ \partial\mathrm{b}\ L_1 &= \partial\mathrm{b}(\mathrm{a}^*(\mathrm{b.a}^*|\lambda)) = \emptyset.(\mathrm{b.a}^*|\lambda)|\lambda.\partial\mathrm{b}(\mathrm{b.a}^*|\lambda) = \\ &\quad \mathrm{a}^* = L_3 \\ \partial\mathrm{a}\ L_2 &= \partial\mathrm{a}(\mathrm{a}^+) = \mathrm{a}^* = L_3 \\ \partial\mathrm{b}\ L_2 &= \partial\mathrm{b}(\mathrm{a}^+) = \emptyset = L_T \\ \partial\mathrm{a}\ L_3 &= \partial\mathrm{a}(\mathrm{a}^*) = \mathrm{a}^* = L_3 \\ \partial\mathrm{b}\ L_3 &= \partial\mathrm{b}(\mathrm{a}^*) = \emptyset = L_T \\ \partial\mathrm{a}\ L_T &= \partial\mathrm{a}\ \emptyset = \emptyset \\ \partial\mathrm{b}\ L_T &= \partial\mathrm{b}\ \emptyset = \emptyset \end{aligned}$$

Con estas derivadas podemos construir el autómata correspondiente. Serán estados finales aquellos que estén representados por expresiones regulares que contengan la cadena nula. En este caso son finales L_1 y L_3. El estado trampa corresponde a la expresión regular $\emptyset$.