

Lenguajes Regulares

Antonio Falcó

Cadenas o palabras I

- Una *cadena o palabra* es una sucesión finita de símbolos.
 - ◆ cadena $\{c, a, d, e, n\}$.
 - ◆ $10001 \{0, 1\}$
- El conjunto de símbolos que empleamos para construir las cadenas o palabras se le llama *alfabeto* y lo denotaremos genéricamente por Σ .
 - ◆ $\Sigma = \{a, b, \dots, x, y, z\}$ forma el alfabeto latino.
 - ◆ $\Sigma = \{0, 1, 2, \dots, 9\}$ forman el alfabeto numérico arábigo.
 - ◆ $\Sigma = \{0, 1\}$ forman el alfabeto binario.
- Una cadena o palabra sobre el alfabeto binario se le llama *cadena binaria*.
- Se dice entonces que $\underline{x}$ es una cadena basada en el alfabeto Σ si

$$\underline{x} = x_1 x_2 \cdots x_n$$

donde $x_i \in \Sigma$ para $i = 1, 2, \dots, n$.

Cadenas o palabras II

- Si $\underline{x}$ no contiene ningún símbolo del alfabeto, entonces diremos que $\underline{x}$ es la *cadena o palabra vacía* y la denotaremos por $\underline{\varepsilon}$.
- Denotaremos la *longitud de una cadena* como

$$|\underline{x}| = \begin{cases} n & \text{si } \underline{x} = x_1 x_2 \cdots x_n \\ 0 & \text{si } \underline{x} = \underline{\varepsilon}. \end{cases}$$

- Sean $\underline{x} = x_1 \cdots x_n$ e $\underline{y} = y_1 \cdots y_m$. Diremos que $\underline{x} = \underline{y}$ si
 1. $n = m$ y
 2. $x_i = y_i$ para $i = 1, 2, \dots, n$.
- Definimos la operación de *concatenación* como

$$\underline{x} \underline{y} = x_1 \cdots x_n y_1 \cdots y_m$$

en este caso

$$|\underline{x} \underline{y}| = |\underline{x}| + |\underline{y}| = n + m.$$

Concatenación y Multiplicación

1. Existe un elemento neutro para la concatenación: la palabra vacía.

$$\underline{x} \underline{\varepsilon} = \underline{\varepsilon} \underline{x} = \underline{x}.$$

2. Se cumple la propiedad asociativa de forma trivial

$$(\underline{x} \underline{y}) \underline{z} = \underline{x} (\underline{y} \underline{z}) = \underline{x} \underline{y} \underline{z}$$

3. No se cumple la propiedad conmutativa

$$01\ 10 \neq 10\ 01$$

4. Existe el equivalente a la “matriz traspuesta”, la llamada *inversión de una cadena o palabra*:

$$\underline{x}^R = (x_1 x_2 \cdots x_{n-1} x_n)^R = x_n x_{n-1} \cdots x_2 x_1.$$

5. Si $\underline{x}^R = \underline{x}$ se dice que la palabra es un *palíndromo*:

$$(01\ 10)^R = 01\ 10.$$

Prefijos, sufijos y potencias

- Si $\underline{x} = \underline{y} \underline{z}$ diremos que $\underline{y}$ es un *prefijo* de la palabra $\underline{x}$ y $\underline{z}$ es un sufijo de la misma palabra.

- Podemos escribir

$$\underline{x} \underline{x} = \underline{x}^2$$

teniendo en cuenta que

$$\underline{x}^1 = \underline{x}.$$

- En general

$$\underline{x}^n = \overbrace{\underline{x} \cdots \underline{x}}^{n-\text{veces}}$$

- De forma evidente

$$\underline{x}^0 = \underline{\varepsilon}.$$

Ejemplos en $\Sigma = \{0, 1\}$

- $1^4 = 1111$.
- $(10)^3 = 10\ 10\ 10$.
- $((10)^3)^R = (10\ 10\ 10)^R = 010101$.
- Resuelve la ecuación $\underline{x}\ 011 = 011\ \underline{x}$.
 - ◆ Es claro que $\underline{x} = \underline{\varepsilon}$ resuelve el problema y que $\underline{x} = 011$ tambien.
 - ◆ Veamos que ocurre si consideramos soluciones del tipo $\underline{x} = 011\underline{y}$. Entonces

$$011\underline{y}\ 011 = 011\ 011\underline{y} \Rightarrow \underline{y}\ 011 = 011\underline{y},$$

el caso anterior nos dice que $\underline{y} = \underline{\varepsilon}, 011 \Rightarrow \underline{x} = 011, (011)^2$

- ◆ Veamos que ocurre si consideramos soluciones del tipo $\underline{x} = \underline{y}011$. Entonces

$$\underline{y}\ 011\ 011 = 011\ \underline{y}\ 011 \Rightarrow \underline{y}\ 011 = 011\underline{y}.$$

En general, $\underline{x} = (011)^n, n \geq 0$.

- Un *lenguaje* sobre un alfabeto Σ es cualquier conjunto formado por cadenas o palabras sobre ese alfabeto.
 - ◆ $A = \{\varepsilon, 0, 1, 00, 11, 01, 10\}$ es un lenguaje sobre $\Sigma = \{0, 1\}$.
 - ◆ A formado por las palabras palíndromos castellanos es un lenguaje sobre el alfabeto latino.
- Si A y B son lenguajes entonces

$$A \cup B = \{\underline{x} : \underline{x} \in A \text{ o } \underline{x} \in B\}$$

$$A \cap B = \{\underline{x} : \underline{x} \in A \text{ y } \underline{x} \in B\}$$

$$A \setminus B = \{\underline{x} : \underline{x} \in A \text{ y } \underline{x} \notin B\}$$

$$A \cdot B = \{\underline{x} : \underline{x} = \underline{y}\underline{z}, \underline{y} \in A \text{ y } \underline{z} \in B\}.$$

$$A^n = \overbrace{A \cdot A \cdot \dots \cdot A}^{n-\text{veces}}, \quad A^0 = \{\varepsilon\}, \quad A^1 = A.$$

Ejemplos

$A = \{0, 1, 00, 10\}$, y $B = \{\underline{\epsilon}, 00, 11\}$.

- $A \cup B = \{0, 1, 00, 10, \underline{\epsilon}, 11\}$

- $A \cap B = \{00\}$

- $A \setminus B = \{0, 1, 10\}$.

- $A \cdot B = \{0, 000, 011, 1, 100, 111, 00, 0000, 0011, 10, 1000, 1011\}$

- $A^2 =$

$\{00, 01, 000, 010, 10, 11, 100, 110, 001, 0000, 0010, 101, 1000, 1010\}$

Puede ocurrir que

$$\{0, 1, 10\} \cap \{\underline{\epsilon}, 00, 11\} = \emptyset,$$

Entonces a $\emptyset$ se le llama *lenguaje vacío* que es aquel que no contiene ninguna palabra o cadena. No confundir con la palabra vacía $\underline{\epsilon}$.

Clausura de Kleene

Dado un lenguaje A definimos su clausura de Kleene A^* como:

$$A^* = A^0 \cup A^1 \cup A^2 \cup \dots \cup A^n \cup \dots = \bigcup_{n=0}^{\infty} A^n.$$

En particular

$$\{0, 1\}^* = \{\varepsilon\} \cup \{0, 1\} \cup \{00, 01, 10, 11\} \cup \dots$$

lo forman todas las cadenas binarias posibles (de cualquier longitud) incluida la palabra vacía.

En consecuencia cualquier lenguaje binario está contenido en la clausura de Kleene del alfabeto binario, esto es, en Σ^* siendo $\Sigma = \{0, 1\}$.

- Si A es un lenguaje en el alfabeto Σ entonces A está contenido en Σ^* .

Lenguaje complementario

Sea A un lenguaje sobre el alfabeto Σ .

- $\overline{A} = \Sigma^* \setminus A = \{\underline{x} \in \Sigma^* : \underline{x} \notin A\}$

- Ejemplo: Si

$$\Sigma = \{0, 1\}$$

entonces

$$A = \{1\} \cdot \Sigma^* = \{1 \underline{x} : \underline{x} \in \Sigma^*\},$$

está formado por todas las palabras que empiezan por 1. En consecuencia $\overline{A}$ estará formado por todas las cadenas que **no** empiezan por 1, y lo podemos escribir como

$$\overline{A} = (\{0\} \cdot \Sigma^*) \cup \{\underline{\varepsilon}\}.$$

Ejemplo

Describe el lenguaje

$$\{0, 10\}^*$$

- Aparte de la palabra vacía y el mismo nos encontramos con

$$\{0, 10\} \cup \{00, 010, 100, 1010\} \cup \dots$$

Nos damos cuenta que siempre acaban en cero y que nunca contiene dos unos seguidos. Nótese que si no contiene unos entonces es de la forma 0^n con $n \geq 0$. Las palabras de una longitud dada que contienen mayor número de unos son de la forma $(10)^n$.

Luego son palabras que nunca contienen la subcadena 11.

Ejemplo

Demuestra que dados dos lenguajes A y B se tiene
 $(A \cdot B)^R = B^R \cdot A^R$

■ En este caso

$$(A \cdot B)^R = \{(\underline{y} \underline{z})^R, \underline{y} \in A \text{ y } \underline{z} \in B\}$$

y

$$B^R \cdot A^R = \{\underline{z}^R \underline{y}^R, \underline{y} \in A \text{ y } \underline{z} \in B\}$$

■ Hay que demostrar que para cualquier par de palabras $\underline{y}, \underline{z}$ se cumple que:

$$(\underline{y} \underline{z})^R = \underline{z}^R \underline{y}^R$$

■ Ahora

$$\begin{aligned} & (y_1 y_2 \cdots y_{n-1} y_n z_1 z_2 \cdots z_{m-1} z_m)^R \\ &= z_m z_{m-1} \cdots z_2 z_1 y_n y_{n-1} y_2 y_1 \\ &= (z_m z_{m-1} \cdots z_2 z_1) (y_n y_{n-1} \cdots y_2 y_1) \\ &= (z_1 z_2 \cdots z_{m-1} z_m)^R (y_1 y_2 \cdots y_{n-1} y_n)^R \end{aligned}$$

Lema de Arden

Lema 1 Sean A y B dos lenguajes sobre un alfabeto Σ , de forma que A no contiene la palabra vacía. Si X es un lenguaje que cumple la relación

$$X = AX \cup B,$$

Entonces $X = A^* \cdot B$.

Ecuaciones con lenguajes

Ejemplo 2 Si U y V son lenguajes sobre $\{a, b\}$ y cumplen las ecuaciones

$$U = \{\varepsilon\} \cup \{a\} \cdot U \cup \{b\} \cdot V \quad (1)$$

$$V = \{\varepsilon\} \cup \{b\} \cdot V \quad (2)$$

Encuentra una representación simple para U y V .

De (2) con $X = V$, $A = \{b\}$ y $B = \{\varepsilon\}$ obtenemos que

$$V = \{b\}^* \cdot \{\varepsilon\} = \{b\}^*$$

Sustituyendo en (1) obtenemos

$$U = \{\varepsilon\} \cup \{a\} \cdot U \cup \{b\} \cdot \{b\}^*, \Rightarrow U = \overbrace{\{\varepsilon\} \cup \{b\} \cdot \{b\}^*}^B \cup \overbrace{\{a\}}^A \cdot U,$$

empleando de nuevo el Lema de Arden obtenemos

$$U = \{a\}^* \cdot (\{\varepsilon\} \cup \{b\} \cdot \{b\}^*) = \{a\}^* \{b\}^*.$$

Una forma de escribir las soluciones del problema anterior sería escribir

$$V = \underline{b}^* \cup \underline{\varepsilon} = \underline{b}^*$$

y

$$U = \underline{a}^* (\underline{\varepsilon} \cup \underline{b} \underline{b}^*) = \underline{a}^* \underline{b}^*.$$

Llegamos a expresiones del tipo

$$A = \underline{aabb}(\underline{ab}^* \cup \underline{b})\underline{bb}^*,$$

esto nos dice que las palabras de A son de la forma:

$$aabb(\underline{\varepsilon} \cup \underline{b})bb = aabb \ bb \cup aabb \ b \ bb$$

Esto es el equivalente al comando de búsqueda
`dir pepet*.*`

identificamos todos los ficheros de estas características con las palabras de un lenguaje descrito mediante lo que se conoce como una *expresión regular*.

Expresiones Regulares

Dado un alfabeto Σ definimos las expresiones regulares sobre el mismo de la forma siguiente:

1. $\emptyset$ es una expresión regular que representa al lenguaje vacío.
2. ε es una expresión regular que representa al lenguaje $\{\varepsilon\}$.
3. Cada elemento a del alfabeto es una expresión regular que representa al lenguaje $\{a\}$.
4. Si r_A y r_B son expresiones regulares que representan a los lenguajes A y B , respectivamente, entonces
 - a) $r_A + r_B$ que representa a $A \cup B$
 - b) $r_A r_B$ que representa a $A \cdot B$ y
 - c) r_A^* que representa a A^*son expresiones regulares.
5. No existen más expresiones regulares sobre Σ que no sea una de estas.

Ejemplos



$$\{a, b\}^* = (a + b)^*$$

que representa

$$(\{a\} \cup \{b\})^*.$$



$$\underline{a}^*(\underline{\varepsilon} \cup \underline{b}\underline{b}^*) = a^*(\varepsilon + bb^*) = a\varepsilon + abb^* = a + abb^*.$$

- Describir en castellano el lenguaje que representa la expresión regular siguiente:

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

En este caso cualquier palabra con el alfabeto $\{a, b\}$ que empiece por la letra a .

Preferencias en las operaciones

1. La clausura de Kleene tiene preferencia sobre la unión y la concatenación.
2. La concatenación tiene preferencia sobre la unión.
- 3.

$$r(s + t) = rs + st$$

4.

$$(r + s)t = rt + st$$

Ejercicio 3 Demuestra que $a^*(a + b)^* = (a + ba^*)^*$.

Definición 4 Denotaremos por $r^+ = r r^*$.

Lenguajes regulares

Definición 5 Diremos que un lenguaje L sobre un alfabeto Σ es regular si podemos construir una expresión regular r de forma que $L = r$.

Ejemplo 6 $\{0, 1\}^*$ es un lenguaje regular ya que

$$\{0, 1\}^* = (0 + 1)^*.$$

Ejemplo 7 Si L es un lenguaje regular sobre un alfabeto Σ . Demostrar que el lenguaje

$$L' = \{w : u w \text{ está en } L \text{ para alguna cadena } u\}$$

es regular.

Lema de Arden (revisitado)

Lema 8 Sean r y s dos expresiones regulares sobre un alfabeto Σ , de forma que r no contiene la palabra vacía. Si x es una expresión regular que cumple la relación

$$x = r x + s,$$

Entonces $x = r^* s$.

Demostración. Hay que demostrar que $x \subset r^* s$ y que $r^* s \subset x$. Para ello empleamos inducción sobre la longitud de $|w|$ en x . Si $w = \varepsilon$, entonces como r no contiene ε se tiene que $w = \varepsilon \in rx + s$ luego $w \in s \subset r^* s$. Supongamos que para cualquier w en x de longitud n se cumple que $w \in r^* s$ y consideremos ahora una cadena w de longitud $n + 1$ como $w \in rx + s$ si $w \in s \subset r^* s$ la propiedad se cumple, en caso contrario tenemos $w \in rx$, luego $w = rw'$ con w' de longitud menor o igual a n . La hipótesis de inducción nos dice que $w' \in r^* s$, luego $w = rw' \in rr^* s \subset r^* s$. ■

Lema de Arden (revisitado)

Lema 9 Sean r y s dos expresiones regulares sobre un alfabeto Σ , de forma que r no contiene la palabra vacía. Si x es una expresión regular que cumple la relación

$$x = r x + s,$$

Entonces $x = r^* s$.

Demostración. Falta demostrar que $r^* s \subset x$. Demostraremos por inducción que $r^n s \subset x$ para todo $n \geq 0$. Para $n = 0$ se tiene que probar que $s \subset x$, pero esto es evidente a partir de la expresión $x = r x + s$. Supongamos que se cumple $r^k s \subset x$ para $k \leq n$, en particular $r^{n+1} s = r(r^n s) \subset r x$. Luego $r^{n+1} s \subset r x \subset r x + s = x$. ■

Principio de inducción generalizado

Deseamos demostrar que los lenguaje regulares cumplen una propiedad Q

1. Demostrar que $\emptyset$ cumple la propiedad Q .
2. Demostrar que para cada a del alfabeto se cumple la propiedad Q
3. Demostrar que si A y B cumplen la propiedad Q entonces $A \cup B$, $A \cap B$ y A^* cumplen la propiedad Q .

Grafos y expresiones regulares

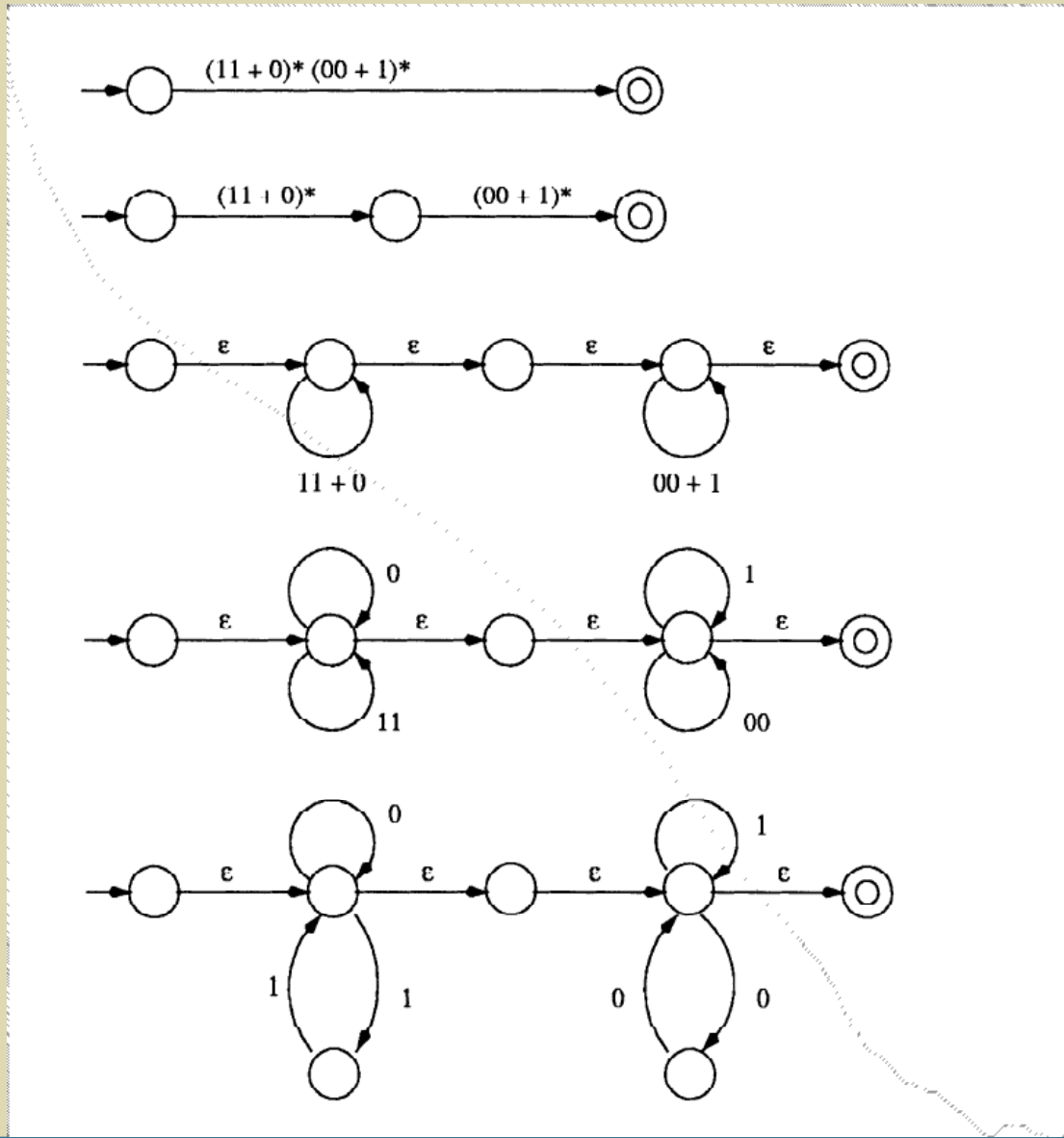
Nos gustaría representar una expresión regular cualquiera r empleando un grafo dirigido y etiquetado.

Ejemplo 10 *Consideremos la expresión regular*

$$r = (11 + 0)^*(00 + 1)^*$$

Construyamos un grafo dirigido $G(r)$ de forma que cualquier cadena x del lenguaje representado por esta expresión regular y que denotaremos $L((11 + 0)^(00 + 1)^*)$ esté completamente caracterizada por los caminos del grafo.*

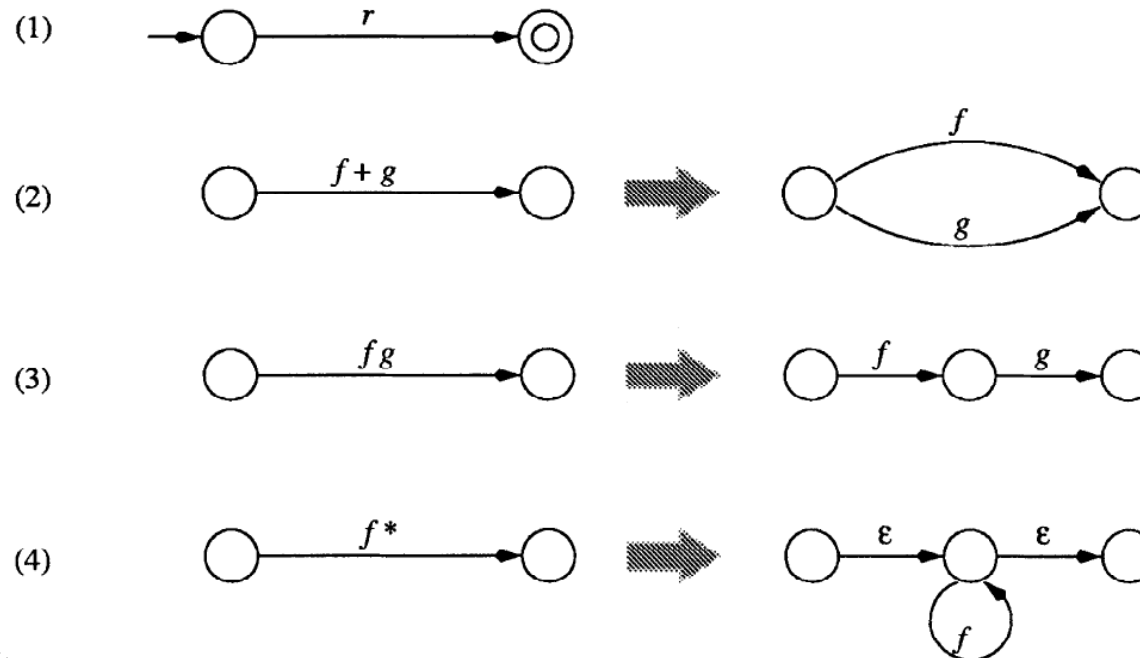
Ejemplo



Construcción del grafo $G(r)$ asociado a r

18

REGULAR LANGUAGES



Equivalencia entre expresiones regulares y grafos

Teorema 11 *Sea r una expresión regular. Una cadena o palabra x pertenece al lenguaje $L(r)$ si y solo si existe un camino en el grafo $G(r)$ desde el vértice inicial al vértice final cuya cadena asociada es x .*

Demostración. Consultar el manual. ■

- En un grafo $G(r)$ una flecha etiquetada con ε se le denomina ε -flecha.
- Sería deseable el poder construir un grafo con el mínimo número de ε -flechas posibles y que mantuviera las propiedades del grafo original.
- Se puede establecer una regla que nos permita eliminar ε -flechas redundantes.

Teorema 12 *Sea r una expresión regular. Entonces una ε -flecha (u, v) en $G(r)$ que es una única flecha de salida desde un vértice no final u o una única flecha de entrada a un vértice no trivial v puede ser colapsado a un vértice simple, preservando las propiedades del anterior Teorema. Si uno de los puntos finales de la ε -flecha es el vértice inicial o el final, entonces lo es a su vez el vértice resultante de la operación de colapso.*

Ejemplo

