Expresiones regulares

Elvira Mayordomo

Universidad de Zaragoza

26 de septiembre de 2013

Contenido de este tema

- Introducción a las expresiones regulares
- Definición de expresión regular
- Ejemplos de expresiones regulares
- Expresiones regulares en Flex
- Analizadores lexicográficos

Expresiones regulares

- Estamos acostumbrados a manejar **expresiones aritméticas**:

$$(5+3)*4$$

$$4*(34+5+2)$$

- Aparecen algunos operadores aritméticos $(+, *)$
- Los operadores tienen precedencias ($*$ se ejecuta antes que $+$ a no ser que un paréntesis diga lo contrario)
- Normalmente usamos nada para la multiplicación $((5+3)4$ y $4(34+5+2))$.
- Cada expresión aritmética representa un número

Expresiones regulares

- Las expresiones regulares operan con lenguajes, utilizando las operaciones unión, concatenación y estrella de Kleene
- Cada expresión regular representa un lenguaje
- Usamos $+$ para la unión (y $\cdot$ o nada para la concatenación)
- La expresión regular a representa el lenguaje $\{a\}$

Ejemplos

- a^*
- $(a+b)$
- $(a+b)a^*$
- 01^*

Aplicaciones

Las expresiones regulares se usan en compilación (por ejemplo, en los analizadores léxicos), procesamiento de lenguaje natural, búsquedas de cadenas en UNIX y muchos editores de texto (Word, vi, Pearl, Emacs, grep), etc, etc.

Definición de expresión regular

Sea Σ un alfabeto. Las expresiones regulares son las siguientes:

1. a (donde $a \in \Sigma$)
2. ϵ
3. $\emptyset$
4. $r_1 + r_2$ (donde r_1 y r_2 son expresiones regulares)
5. $r_1 \cdot r_2$ (donde r_1 y r_2 son expresiones regulares)
6. r^* (donde r es una expresión regular).

Es una definición recursiva, una expresión regular se construye a partir de 1., 2. y 3. aplicando las reglas 4., 5. y 6.

Ejemplos

Son expresiones regulares:

- $(b + a) \cdot ba^*$
- $aa(a + b)^*(ba + \epsilon)$
- $\emptyset \cdot b$

No son expresiones regulares:

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

Definición de expresión regular (2)

Cada expresión regular representa un lenguaje. Para una expresión regular r, $L(r)$ es el lenguaje representado por r definido como sigue:

1. $L(a) = \{a\}$ (donde $a \in \Sigma$)
2. $L(\epsilon) = \{\epsilon\}$
3. $L(\emptyset) = \emptyset$
4. $L(r_1 + r_2) = L(r_1) \cup L(r_2)$ (donde r_1 y r_2 son expresiones regulares)
5. $L(r_1 \cdot r_2) = L(r_1) \cdot L(r_2)$ (donde r_1 y r_2 son expresiones regulares)
6. $L(r^*) = L(r)^*$ (donde r es una expresión regular).

Es una definición recursiva, el lenguaje representado por una expresión regular se construye a partir de 1., 2. y 3. aplicando las reglas 4., 5. y 6.

Ejemplos

- $L(a^*) = \{a\}^* = \{a^n \mid n \in \mathbb{N}\}$
- $L(a+b) = \{a, b\}$
- $L((a+b)a^*) = \{a, b\} \cdot \{a\}^* = \{a^n \mid n \geq 1\} \cup \{ba^n \mid n \in \mathbb{N}\}$
- $L(01^*) = \{0\} \cdot \{1\}^* = \{01^n \mid n \in \mathbb{N}\}$
- $L((1+0) \cdot 10^*) = \{0,1\} \cdot \{1\} \cdot \{0\}^* = \{010^n \mid n \in \mathbb{N}\} \cup \{110^n \mid n \in \mathbb{N}\}$
- $L(\emptyset \cdot b) = \emptyset \cdot \{b\} = \emptyset$

Expresiones regulares equivalentes

- Son las que representan el mismo lenguaje

$(a+b)a^*$
$aa^* + ba^*$
Decimos $(a+b)a^* = aa^* + ba^*$

$(a+ba)^*$??
$(a^*(ba)^*)^*$??

Propiedades

Dadas dos expresiones regulares r y s se cumplen entre otras:

- $r + s = s + r$
- $r + \emptyset = r$
- $r + r = r$
- $r\epsilon = \epsilon r = r$
- $r(sr)^* = (rs)^* r$

$\ldots$

No es necesario recordarlas, es general se siguen de la definición y veremos métodos para comprobar si dos expresiones regulares son equivalentes.
Tenéis una lista de propiedades en la página 51 del Kelley.
a partir de ahora, abreviamos expresión regular como e.r.

Expresiones regulares en Flex

Hay muchas herramientas informáticas que usan e.r.(UNIX, Word, Flex, Emacs, Perl, ...), cada una cambia ligeramente la forma de escribirlas. Por ejemplo este cuadro muestra algunas correspondencias entre la notación que usaremos en las clases y la que se utiliza en *Flex* (clases de laboratorio).

Operación	**e.r. Teoría**	**e.r. Flex**
Concatena	$\cdot$ ó nada	nada
Unión	$+$	$\vert$
Kleene	*	$*$
$+$	$^+$	$+$
Paréntesis	( )	( )
$a, b, c, 0, 1, 2$	$(a+b+c+0+1+2)$	$[abc012]$ ó $[a-c0-2]$
cadena vacía ó a	$\epsilon + a$	$a?$
Cadena vacía	ϵ	Sin equivalencia (ver $*$ ó ?)
Conjunto vacío	$\emptyset$	Sin equivalencia

Analizadores lexicográficos

- Un analizador lexicográfico es un programa que recorre un texto buscando fragmentos que cumplan un patrón y realiza acciones con estos fragmentos.
- Los fragmentos que se buscan se expresan con e.r.
- Por ejemplo, un programa que busca todas las apariciones de varios espacios seguidos y las cambia por uno solo.
- En prácticas usaréis Flex, un creador de analizadores lexicográficos.

Bibliografía

- Sipser (2a edición), páginas 63 a 66 (en sección 1.3) .
- Kelley, sección 2.2.
- Sobre analizadores lexicográficos y flex: guiones y manual de flex en prácticas.