Compiladores: Sesión 3. Análisis léxico, expresiones regulares

Prof. Gloria Inés Alvarez V.

Departamento de Ciencias e Ingeniería de la Computación
Pontificia Universidad Javeriana Cali

29 de enero de 2008

Definiciones básicas 1

Definición

Un alfabeto es un conjunto finito y no vacío de símbolos

Definición

Una cadena (ó palabra) es una secuencia finita de símbolos de un alfabeto. La cadena vacía es aquella que contiene cero símbolos y se denota ε*. La longitud de una cadena es el número de símbolos que contiene y se denota* $|abb| = 3$

Definiciones básicas 2

Definición

Potencias de un alfabeto: dado un alfabeto Σ, *el conjunto* Σ^k *contiene todas las cadenas de longitud* k *que se pueden formar con símbolos del alfabeto*

Notar la diferencia que hay entre Σ y Σ^1

Definición

Se denota Σ^* *al conjunto de todas las cadenas que se pueden generar a partir del alfabeto* Σ. *Es decir,* $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \ldots$. *por su parte,* $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \ldots$. *En otras palabras,* $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$

Definiciones básicas 3

Definición

Dado un alfabeto Σ, L *es un lenguaje sobre* Σ *si y sólo si* $L \subseteq \Sigma^*$

Notar que Σ^*, $\emptyset$ y $\{\varepsilon\}$ son lenguajes que están definidos para cualquier alfabeto. Y notar que $\emptyset \neq \{\varepsilon\}$

Expresiones Regulares

- Son una notación para definir lenguajes
- Se ha demostrado que pueden expresar la clase de los lenguajes regulares
- Permiten hacer una definición algebraica de un lenguaje
- Se expresa el lenguaje de manera declarativa

Aplicaciones de las expresiones regulares

- Búsqueda de archivos en varios comandos de UNIX, por ejemplo grep
- Análisis léxico de un compilador

Operadores de las expresiones regulares

Dadas dos expresiones regulares E_1, E_2, el resultado de aplicar cualquiera de estas tres operaciones, dá como resultado otra expresión regular:

- Unión $(E_1 \cup E_2) = \{x \mid x \in E_1 \text{ ó} x \in E_2 \text{ ó } x \in E_1 \cap E_2\}$.
- Concatenación $E_1 \cdot E_2 = \{xy \mid x \in E_1, y \in E_2\}$
- Clausura o Estrella de Kleene $(E_1^*) = \cup_{i \geq 0} E_1^i$

Lenguaje asociado con una expresión regular

Definición

Casos base: $L(\varepsilon) = \{\varepsilon\}, L(\emptyset) = \emptyset$*, si* $a \in \Sigma$*,* $L(a) = \{a\}$ *Una variable, normalmente representada con una letra mayúscula,por ejemplo L, representa un lenguaje.*
Caso general:

- *Si E y F son expresiones regulares,* $L(E + F) = L(E) \cup L(F)$
- *Si E y F son expresiones regulares,* $L(EF) = L(E)L(F)$
- *Si E y F son expresiones regulares,* $L(E^*) = (L(E))^*$
- *Si E y F son expresiones regulares,* $L((E)) = L(E)$

Orden de precedencia de los operadores

Alta	Estrella
	Punto
Baja	Unión

El operador de más alta precedencia elige primero sus argumentos.
Ejemplo: $01^* + 1$ se interpreta como: $(0(1^*)) + 1$

Autómata de Estados Finito

Definición

Un autómata es una 5-tupla de la forma: $\mathcal{A} = \{Q, \Sigma, \delta, I, F\}$ *donde:*

- Q *es un conjunto de estados.*
- Σ *es un alfabeto de símbolos*
- δ *es una función* $Q \times \Sigma \rightarrow 2^Q$
- I *es un conjunto de estados iniciales que pertenecen a* Q
- F *es un conjunto de estados finales, de aceptación, que pertenecen a* Q

Definición

Si $|I| = 1$ *y* $\forall q \in Q, a \in \Sigma |\delta(q, a)| \leq 1$*, se dice que el autómata es determinista, en caso contrario, se dice que es no determinista.*

Lenguaje asociado a un Autómata de Estados Finito

Definición

Sea $\mathcal{A} = \{Q, \Sigma, \delta, I, F\}$*, el lenguaje asociado a* $\mathcal{A}$ *es:*
$L(\mathcal{A}) = \{x \mid \exists q_0 \in I, \delta(q_0, x) \cap F \neq \emptyset\}$

Bibliografía

Introduction to Automata theory, languages and computation. Hopcroft J., Motwani R., Ullman J. Segunda Edicion. Adisson Wesley. 2001.