Expresiones Regulares

Horst H. von Brand
vonbrand@inf.utfsm.cl

Departamento de Informática
Universidad Técnica Federico Santa María

Contenido

Introducción

Hablamos de *lenguajes*, compuestos de *palabras*, a su vez compuestas de *símbolos*. Definimos operaciones entre lenguajes, que aprovecharemos para definir nuevos lenguajes sistemáticamente, mediante *expresiones regulares*.

Expresiones regulares

Definición

Una *expresión regular* sobre el alfabeto Σ se define junto al lenguaje que denota mediante:

i) La expresión regular $\varnothing$ denota el lenguaje $\varnothing$.

ii) La expresión regular ε denota el lenguaje $\{\varepsilon\}$.

iii) Para cada $a \in \Sigma$ la expresión regular a denota el lenguaje $\{a\}$.

Sean R y S expresiones regulares, que denotan $\mathscr{L}(R)$ y $\mathscr{L}(S)$, respectivamente.

iv) La expresión regular $(R)\,|\,(S)$ denota $\mathscr{L}(R) \cup \mathscr{L}(S)$.

v) La expresión regular $(R)\cdot(S)$ denota $\mathscr{L}(R)\cdot\mathscr{L}(S)$.

vi) La expresión regular $(R)^*$ denota $(\mathscr{L}(R))^*$.

Expresiones regulares

Algunas extensiones extraoficiales, usadas más que nada para simplificar:

vii) La expresión $(R)^n$, para $n \geq 0$ fijo, denota $(\mathscr{L}(R))^n$.

viii) La expresión $(R)^+$ denota $(\mathscr{L}(R))^+$.

Para evitar interminables paréntesis y signos de multiplicación, nuevamente usamos las convenciones tomadas del álgebra: primero potencias (estrella, más, potencias), luego multiplicaciones (concatenación) finalmente sumas (unión); comúnmente omitimos el signo de multiplicación.
Para el alfabeto Σ usaremos Σ^* (que en rigor es un crimen, Σ no es un lenguaje) para denotar al lenguaje de todas las palabras que se pueden formar con símbolos de Σ.

Regex

Cuidado, muchas herramientas y lenguajes de programación manejan patrones de búsqueda a los que llaman «expresiones regulares» (en inglés, *regular expressions*). Tienen solo un parecido general con la noción matemática (fueron inspirados por ella, claro). Suelen incluir operaciones adicionales por comodidad, o tienen restricciones importantes. Algunos sistemas los llaman *regex*, para distinguir.
Buscar «expresiones regulares» en la red generalmente da con páginas que discuten regex, más que nada distinguiendo los múltiples dialectos incompatibles.

Ejemplos

Supongamos el alfabeto $\Sigma = \{a, b, c\}$ en lo que sigue.
Nuestras convenciones indican:

$((a) \cdot (b)^*)$ se escribe ab^*
$(((a) \mid (b)) \mid ((c)^*)$ se escribe $a \mid b \mid c^*$
$(((a) \cdot (b)) \cdot ((c)^*)$ se escribe abc^*

Compare:

ab^* denota $\{a, ab, abb, \ldots\}$
$(ab)^*$ denota $\{\varepsilon, ab, abab, ababab, \ldots\}$
$a \mid b^*$ denota $\{\varepsilon, a, b, bb, \ldots\}$
$(a \mid b)^*$ denota $\{\varepsilon, a, b, aa, ab, ba, bb, aaa, \ldots\}$

Ejemplos

Supongamos el alfabeto $\Sigma = \{a, b\}$ en lo que sigue.

a^*	denota	$\{\varepsilon, a, aa, aaa, \ldots\}$
ab^*	denota	$\{a, ab, abb, \ldots\}$
ab^2	denota	$\{abb\}$
$(a\|b)^3$	denota	$\{aaa, aab, aba, \ldots, bbb\}$
$(ab)^3$	denota	$\{ababab\}$
$a \mid b^*$	denota	$\{\varepsilon, a, b, bb, \ldots\}$

Ejemplos

$(a \mid b)^*$ denota $\{\varepsilon, a, b, aa, ab, ba, bb, aaa, \ldots\}$

$(b \mid ab)^*(\varepsilon \mid a)$ denota palabras sin aa

$(ab^+ \mid ba^+)^*$ denota ...

Ya las sencillas expresiones regulares finales muestra que son capaces de representar lenguajes nada triviales.

Simplificaciones

Estamos describiendo operaciones entre lenguajes, todo lo que hemos dado como identidades es aplicable.
Básicamente, podemos aplicar las reglas del álgebra (¡Con cierto cuidado! No todas las propiedades de un campo se aplican.).

Resumen

- Definimos expresiones regulares y los lenguajes que denotan.
- Note que es una definición recursiva, demostraciones con expresiones regulares naturalmente serán por inducción, siguiendo la estructura de la definición.
- Mostramos algunas reglas de simplificación en las transparencias anteriores.

Es importante familiarizarse con expresiones regulares y poder leer los lenguajes denotados. Usaremos esta notación frecuentemente durante el curso.