Expressões Regulares

Prof^a Jerusa Marchi

jerusa@inf.ufsc.br

Departamento de Informática e Estatística

Universidade Federal de Santa Catarina

e-mail: jerusa@inf.ufsc.br

Linguagens Regulares

- Geradas por Gramáticas Regulares
- Reconhecidas por Autômatos Finitos

Linguagens Regulares

- Conjunto de linguagens decidíveis bastante simples e com propriedades bem definidas e compreendidas
  - União
  - Concatenação
  - Fechamento

Linguagens Regulares

- União - $(L \cup M)$
  - conjunto de palavras que pertencem a L e M
    - ex.: $L = \{001, 10, 111\}$ e $M = \{\epsilon, 001\}$

$$L \cup M = \{\epsilon, 10, 001, 111\}$$

- Concatenação $(L \cdot M$ ou $LM)$
  - conjunto de palavras formadas pela concatenação de palavras de L com palavras de M
    - ex.: $L = \{001, 10, 111\}$ e $M = \{\epsilon, 001\}$

$$L \cdot M = \{001, 10, 111, 001001, 10001, 111001\}$$

Linguagens Regulares

- Fechamento (L^*)
  - conjunto de palavras formadas pegando qualquer número de palavras de L concatenadas.
    - ex.: $L = \{0, 1\}$ então L^* contém todas as palavras formadas por 0's e 1's

      $L = \{0, 11\}$ então L^* consiste naquelas palavras formadas por 0's e 1's tal que 1 venha sempre em pares.

      $\{011, 11110, \epsilon\}$ pertencem a L

      $\{01011, 101\}$ não pertencem a L

Linguagens Regulares

- Formalmente:
  - Fechamento:

$$L^{\star} = \bigcup_{i=0}^{\infty} L^i = L^0 \cup L^1 \cup L^2 \cup \ldots$$

  - Fechamento Positivo:

$$L^{+} = \bigcup_{i=1}^{\infty} L^i = L^1 \cup L^2 \cup \ldots$$

Linguagens Regulares

- Seja L o conjunto de letras $\{A, B, \cdots, Z, a, b, \cdots, z\}$ e seja D o conjunto de dígitos $\{0, 1, \cdots, 9\}$.
- Outras linguagens que podem ser construídas a partir de L e D:
  1. $L \cup D$ o conjunto de letras e dígitos - 62 palavras, cada uma sendo uma letra ou um dígito
  2. LD é o conjunto de 520 palavras de tamanho 2, cada uma consistindo de uma letra seguida por um dígito
  3. L^4 é o conjunto de todas as palavras de 4 letras
  4. L^* é o conjunto de todas as palavras, incluindo a sentença vazia
  5. $L(L \cup D)^*$ é o conjunto de todas as palavras de letras e dígitos iniciadas com uma letra
  6. D^+ é o conjunto de todas as palavras formadas por um ou mais dígitos

Expressões Regulares

- Uma Linguagem Regular pode ser descrita por expressões simples, chamadas **Expressões Regulares** (ER)
- Uma expressão regular é construída recursivamente a partir de expressões regulares mais simples, usando duas regras básicas e indução

Expressões Regulares: Definição

- Base:
  - ε é uma expressão regular, e $L(\varepsilon)$ é $\{\varepsilon\}$, que a linguagem que tem somente uma palavra que é a palavra vazia
  - Se a é um símbolo em Σ, então a é uma expressão regular, e $L(a) = \{a\}$, que é a linguagem com somente uma palavra, de tamanho um, com a nesta posição
- Cada expressão regular r denota a linguagem $L(r)$, que é definida recursivamente a partir das linguagens denotadas pelas subexpressões de r

Expressões Regulares: Definição

- Indução:
  - Se (r) e (s) são expressões regulares, então $r \mid s$ ou $r + s$ é uma expressão regular que denota a união de $L(r)$ e $L(s)$. Isto é $L(r + s) = L(r) \cup L(s)$.
  - Se (r) e (s) são expressões regulares, então rs é uma expressão regular que denota a concatenação de $L(r)$ e $L(s)$. Isto é $L(rs) = L(r)L(s)$.
  - Se (r) é uma expressão regular, então r^* é uma expressão regular que denota o fechamento de $L(r)$. Isto é $L(r^*) = (L(r))^*$.
  - Se (r) é uma expressão regular, então (r) é também uma expressão regular, que denota a mesma linguagem de r. Formalmente $L((r)) = L(r)$.

Precedência de Operadores

- O operador de fechamento ($*$) tem maior precedência
- Seguido da Concatenação (.)
- Por fim, União ($+$)

Exemplos

- $01^* + 1 \rightarrow (0(1)^*) + 1$
- $0 + 11* \rightarrow 0 + (1(1)^*)$
- $(01)^* + 1$
- $0(1^* + 1)$

Exemplos

- Seja $\Sigma = \{a, b\}$

1. A ER $a \mid b$ denota a linguagem $\{a, b\}$
2. $(a \mid b)(a \mid b)$ denota $\{aa, ab, ba, bb\}$, a linguagem de todos as palavras de tamanho 2 sobre Σ
   - $aa \mid ab \mid ba \mid bb$ é outra ER para a mesma linguagem
3. a^* denota a linguagem consistindo de todas as palavras com zero ou mais a's, isto é $\{\varepsilon, a, aa, aaa, \cdots\}$
4. $(a \mid b)^*$ denota o conjunto de todas as palavras consistindo de zero ou mais ocorrências de a ou b, isto é $\{\varepsilon, a, b, aa, ab, ba, bb, aaa, \cdots\}$
   - Outra ER para a mesma linguagem é $(a^*b^*)^*$
5. $a \mid a^*b$ denota a linguagem $\{a, b, ab, aab, aaab, \cdots\}$, isto é, palavras consistindo de zero ou mais a's e acabando com b.

Exemplos

- 00 é uma ER que denota a linguagem $\{00\}$
- $(0+1)^*$ é uma ER que denota todas as cadeias de 0's e 1's
- $(0+1)^*00(0+1)^*$ denota as cadeias de 0's e 1's com pelo menos dois 0's consecutivos
- $(1+10)*$ denota as cadeias de 0's e 1's que começam com 1 e não tem 0's consecutivos
- $(0+1)^*001$ denota as cadeias de 0's e 1's que terminam por 001

Leis algébricas

- Seja r,s e t ER arbitrárias:

Lei	Descrição
$r \mid s = r \mid r$	$\mid$ é comutativo
$r \mid (s \mid t) = (r \mid s) \mid t$	$\mid$ é associativo
$r(st) = (rs)t$	Contatenação é associativa
$r(s \mid t) = rs \mid rt; (s \mid t)r = sr \mid tr$	Concatenação é distributiva
$\varepsilon r = r\varepsilon = r$	ε é a identidade para a concatenação
$r^* = (r \mid \varepsilon)^*$	ε é garantido no fechamento
$r^{**} = r^*$	* é idempotente

Exemplo

- Escrever uma ER para o conjunto de palavras que consistem em 0's e 1's dispostos alternadamente.
  - ER para todas as palavras formadas por 01
    $(01)^* = \{01, 0101, 0101 \cdots 01\}$
  - Ainda faltam as palavras do tipo $\{1 \cdots 0\}$, $\{0 \cdots 0\}$ e $\{1 \cdots 1\}$

ER = $(01)^* + (10)^* + 0(10)^* + 1(01)^*$

ou

ER = $(\epsilon + 1)(01)^*(\epsilon + 0)$

Definições Regulares

- Podemos nomear Expressões Regulares e usar esses nomes em ER subsequentes, como se os nomes fossem os próprios símbolos
- Se Σ é um alfabeto de símbolos básicos, então uma *definição regular* é uma sequência de defnições da forma:

$$
\begin{aligned}
d_1 &\rightarrow r_1 \\
d_2 &\rightarrow r_2 \\
&\cdots \\
d_n &\rightarrow r_n
\end{aligned}
$$

onde:

1. Cada d_i é um novo símbolo, não em Σ e não o mesmo que qualquer outro d_i
2. Cada r_i é uma expressão regular envolvendo símbolos no alfabeto $\Sigma \cup \{d_1, d_2, \cdots, d_n\}$

Exemplo

- Os identificadores da linguagem C são cadeias de letras e dígitos e *underscore*.

$$\begin{array}{lcl} letter\_ & \rightarrow & A \mid B \mid C \mid \cdots \mid Z \mid a \mid b \mid \cdots \mid z \mid \_ \\ digit & \rightarrow & 0 \mid 1 \mid \cdots \mid 9 \\ id & \rightarrow & letter\_(letter\_ \mid digit)^* \end{array}$$

Extensões de ER

- Desde que Kleene introduziu ER com os operadores básicos de união, concatenação e fecho de Kleene, na década de 50, divesas extensões foram incorporadas. São algumas:
  1. *Uma ou mais instâncias* - o operador $+$ denota o fecho positivo de uma ER. Tem a mesma precedência de $*$. Duas leis algébricas úteis são: $r^* = r^+ \mid \varepsilon$ e $r^+ = rr^* = r^*r$
  2. *Zero ou uma instância* - O operador $?$ significa "zero ou uma ocorrência". Ou seja $r^? = r \mid \varepsilon$. Tem a mesma precedência e associatividade de $+$ e $*$
  3. *Classes de caracteres* - uma ER $a_1 \mid a_2 \mid \cdots \mid a_n$ onde $a_i \in \Sigma$, pode ser substituida pela abreviação $[a_1 a_2 \cdots a_n]$. Se for uma sequência lógica, como letras maiúsculas consecutivas, podemos substituir por $a_1 - a_n$.

Exemplo

- Podemos reencrever a ER dos identificadores da linguagem C como:

$$
\begin{array}{lcl}
letter\_ & \rightarrow & [A - Za - z\_] \\
digit & \rightarrow & [0 - 9] \\
id & \rightarrow & letter\_(letter\_ \mid digit)^*
\end{array}
$$