AUTÓMATOS E LINGUAGENS FORMAIS

(LCC/LMAT)

1. Linguagens Formais e Expressões Regulares

Departamento de Matemática

Universidade do Minho

2022/2023

Definição

- Um alfabeto é um conjunto não vazio.
  Habitualmente, notaremos alfabetos por: $A, A', A_0, \ldots$.
- Uma letra de uma alfabeto A é um elemento de A.
  Habitualmente, notaremos letras de alfabetos por: $a, b, c, a', a_0, \ldots$.
- Uma palavra sobre um alfabeto A é uma sequência finita de letras de A, possivelmente vazia.
  Consequentemente, duas palavras são iguais quando as respetivas sequências de letras forem iguais.
  Habitualmente, notaremos palavras por: $u, v, w, x, y, z, u', u_0, \ldots$.
- Chamaremos palavra vazia à sequência vazia de letras, que notaremos por ϵ.
- A notação $a_1 \ldots a_n$, com $n \geq 1$, representará uma palavra não vazia, cuja primeira letra é a_1, a segunda a_2, etc.

Por exemplo,

$$1,\ 01,\ 11,\ 1001,\ 00001,\ \epsilon$$

são palavras sobre o alfabeto $\{0, 1\}$. Por outro lado,

$$aba,\ ab,\ a,\ c,\ bbcac,\ \epsilon$$

são palavras sobre o alfabeto $\{a, b, c\}$.

Definição

- A^+ notará o conjunto das palavras não vazias sobre o alfabeto A, i.e.,

$$A^+ = \{a_1 a_2 \cdots a_n \mid n \in \mathbb{N}, a_1, a_2, \ldots, a_n \in A\}.$$

- A^* notará o conjunto das palavras sobre o alfabeto A, i.e.,

$$A^* = \{\epsilon\} \cup A^+.$$

Por exemplo, sendo $A = \{a, b\}$, tem-se:

$A^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, bab, bba, bbb, \ldots\}$.
Exercício: mostre que, para A finito, o conjunto A^* é numerável.

Observação

Para provar propriedades sobre palavras é, por vezes, útil ter-se uma definição indutiva do conjunto A^*. Como se pode verificar, A^* é o conjunto X definido indutivamente pelas regras:

(i) $\epsilon \in X$;

(ii) Se $u \in X$ e $a \in A$, então $ua \in X$;

uma vez convencionado que $a_1 ... a_n$ abrevia $\epsilon a_1 ... a_n$.

Observação

A anterior caracterização indutiva de A^* não só permite obter um princípio de indução em palavras, como permite também obter um princípio de recursão em palavras.

Definição

O comprimento de uma palavra u é o comprimento da respetiva sequência de letras, sendo notado por $|u|$.

Por exemplo, fixando o alfabeto $\{a, b, c\}$:

$$|\epsilon| = 0, \ |a| = 1, \ |abc| = 3, \ |bbab| = 4.$$

Observação

O comprimento de uma palavra sobre um alfabeto A corresponde a uma operação de A^* em $\mathbb{N}_0$, que pode ser caracterizada por recursão em palavras do seguinte modo:

1. $|\epsilon| = 0$;
2. $|ua| = |u| + 1$, para todo $u \in A^*$ e para todo $a \in A$.

Exercício: verifique as exemplificações acima utilizando esta caracterização recursiva de comprimento de uma palavra.

Definição

O número de ocorrências de uma letra a numa palavra u é notado por $|u|_a$.

Por exemplo, fixando o alfabeto $\{a, b, c\}$:

$$|baba|_c = 0, \ |acaba|_a = 3, \ |acaba|_b = 1.$$

Proposição

Sejam A um alfabeto e $u \in A^*$. Então: $\quad |u| = \sum_{a \in A} |u|_a.$

Exercício: mostre a proposição, recorrendo a indução em palavras ou, em alternativa, a indução no comprimento de palavras .

Definição

A concatenação de uma palavra u com uma palavra v será notada por $u \cdot v$ ou, simplesmente, por uv, sendo dada pela concatenação das respetivas listas. Dito de outro modo:

- se $u = \epsilon$, então $u \cdot v = v$;
- se $u = a_1 \ldots a_n$ e $v = \epsilon$, então $u \cdot v = u$;
- se $u = a_1 \ldots a_n$ e $v = b_1 \ldots b_m$, então $u \cdot v = a_1 \ldots a_n b_1 \ldots b_m$.

Por exemplo, para as palavras $u = abc$ e $v = aa$ (sobre $\{a, b, c\}$),

$$
\begin{aligned}
& u \cdot v = abcaa, \\
& v \cdot u = aaabc, \\
& u \cdot \epsilon = abc, \\
& (v \cdot u) \cdot v = aaabcaa = v \cdot (u \cdot v)
\end{aligned}
$$

Exercício: dê uma definição da operação de concatenação $u \cdot v$ por recursão na palavra v.

Observação

A operação de concatenção de palavras é associativa,com elemento neutro ϵ, pelo que, dado um alfabeto A,

$$(A^*, \cdot) \text{ é um monóide,}$$

chamado o monóide livre gerado por A.

No entanto, a concatenção de palavras não é comutativa.

Proposição

Para $u, v, w \in A^*$, tem-se:

1. $uv = uw \Rightarrow v = w$ (lei do corte à esquerda);
2. $vu = wu \Rightarrow v = w$ (lei do corte à direita);
3. $|uv| = |u| + |v|$ e $|uv|_a = |u|_a + |v|_a$ (para todo $a \in A$).

Definição

Sejam $u \in A^*$ e $n \in \mathbb{N}_0$. A potência-n de u corresponderá à "concatenação de n cópias de u", sendo notada por u^n e definida recursivamente por:

$$u^n = \begin{cases} \epsilon & \text{se } n = 0 \\ u^{n-1}u & \text{se } n \geq 1. \end{cases}$$

Proposição

Para toda a palavra u e para todo $n, m \in \mathbb{N}_0$,

$$u^{n+m} = u^n u^m, \quad (u^n)^m = u^{nm}, \quad |u^n| = n|u|.$$

Definição

Sejam u e v duas palavras de A^*. Diz-se que:

- u é um fator de v quando existem $x, y \in A^*$ tais que $xuy = v$;
- u é um prefixo de v quando existe $y \in A^*$ tal que $uy = v$;
- u é um sufixo de v quando existe $x \in A^*$ tal que $xu = v$.

Por exemplo, sendo $v = baba$:

- os fatores de v são ϵ, b, a, ba, ab, bab, aba, v;
- os prefixos de v são ϵ, b, ba, bab, v;
- os sufixos de v são ϵ, a, ba, aba, v.

Definição

A palavra inversa de uma palavra $u \in A^*$ denota-se por u^I e define-se recursivamente por:

$$u^I = \begin{cases} \epsilon & \text{se } u = \epsilon \\ av^I & \text{se } u = va \quad \text{com } v \in A^* \text{ e } a \in A. \end{cases}$$

Proposição

Para quaisquer $a_1, a_2, \ldots, a_n \in A$ e $u, v \in A^*$,

$$(a_1 a_2 \cdots a_n)^I = a_n \cdots a_2 a_1,$$
$$(uv)^I = v^I u^I,$$
$$(u^I)^I = u.$$

Definição

Uma linguagem sobre um alfabeto A é um subconjunto de A^*.
Habitualmente, notaremos linguagens por: $L, K, M, L', L_0, \ldots$.

São exemplos linguagens sobre $A = \{a, b\}$:

$$\emptyset, \{\epsilon\}, \{a\}, A, \{aa, aba, bbb, ababa\}, \{a^m b^n : m, n \in \mathbb{N}\}, A^+, A^* .$$

Observação

1. O conjunto de todas as linguagens sobre o alfabeto A é $\mathcal{P}(A^*) = \{L : L \subseteq A^*\}$.
2. Para uma alfabeto finito A, dado que A^* é um conjunto infinito (numerável), $\mathcal{P}(A^*)$ é um conjunto infinito não numerável.

Exercício

Defina indutivamente as seguintes linguagens:

a) $L_0 = \{u \in \{0,1\}^* : \ 1 \text{ é fator de } u\}$;

b) $L_1 = \{u \in \{a,b\}^* : |u|_a = |u|_b\}$;

c) $L_2 = \{u \in \{0,1\}^* : \ |u|_0 \text{ é par}\}$;

d) $L_3 = \{w \in \{a,b,c\}^* : w = w^I\}$.

Mostremos, por exemplo, que L_1 é o conjunto L das palavras sobre o alfabeto $A = \{a,b\}$ definido indutivamente pelas regras seguintes:

1. $\epsilon \in L$;
2. Se $w \in L$, então $awb \in L$;
3. Se $w \in L$, então $bwa \in L$;
4. Se $w_1, w_2 \in L$, então $w_1 w_2 \in L$.

Para provar a inclusão $L \subseteq L_1$, precisaremos do Princípio de indução estrutural associado a esta definição indutiva de L.

Princípio de indução estrutural para L

Seja $P(x)$ uma condição sobre $x \in L$.
Se:

1. $P(\epsilon)$;
2. para qualquer $w \in L$, se $P(w)$, então $P(awb)$;
3. para qualquer $w \in L$, se $P(w)$, então $P(bwa)$;
4. para quaisquer $w_1, w_2 \in L$, se $P(w_1)$ e $P(w_2)$, então $P(w_1 w_2)$;

então $P(x)$, para todo $x \in L$.

Mostremos por indução estrutural sobre L que, para cada $x \in L$, $|x|_a = |x|_b$. Para $x \in L$, seja $P(x)$ a condição: $|x|_a = |x|_b$.

1. A propriedade $P(\epsilon)$ é $|\epsilon|_a = |\epsilon|_b$. Ora, como $|\epsilon|_a = 0 = |\epsilon|_b$, tem-se $P(\epsilon)$.
2. Seja $w \in L$ e suponhamos $P(w)$, por hipótese de indução. Ou seja, suponhamos que: $|w|_a = |w|_b$. Então
$$|awb|_a = |w|_a + 1 = |w|_b + 1 = |awb|_b.$$
Provou-se assim $P(awb)$.
3. Esta condição prova-se de forma análoga à anterior.
4. Sejam $w_1, w_2 \in L$ e suponhamos, por H.I., $P(w_1)$ e $P(w_2)$, isto é: $|w_1|_a = |w_1|_b$ e $|w_2|_a = |w_2|_b$. Logo
$$|w_1 w_2|_a = |w_1|_a + |w_2|_a = |w_1|_b + |w_2|_b = |w_1 w_2|_b.$$
Provou-se assim $P(w_1 w_2)$.

Pelo Princípio de indução estrutural para L, de 1 a 4, conclui-se que $P(x)$ é verdadeira para todo o $x \in L$. Provou-se assim que $L \subseteq L_1$.

Mostremos agora a inclusão $L_1 \subseteq L$, ou seja, mostremos que, para cada $x \in L_1$, $x \in L$. Para $x \in L_1$, seja $Q(x)$ a condição $x \in L$.
A prova será feita por indução no comprimento da palavra x.

1. Caso $|x| = 0$. Neste caso $x = \epsilon$. Como, pela regra 1 da definição de L, se tem $\epsilon \in L$, segue $Q(x)$.
2. Caso $|x| > 0$. Suponhamos, por H.I., que $Q(y)$ é verdadeira para todas as palavras $y \in L_1$ tais que $|y| < |x|$. Existem 4 possibilidades:
   - $x = awb$ (ou $x = bwa$) para algum $w \in A^*$. Dado que $x \in L_1$, tem-se $|x|_a = |x|_b$, donde $|w|_a = |w|_b$. Portanto $w \in L_1$ e, pela H.I., $w \in L$. Daqui resulta, pela regra 2 da definição de L, que $x \in L$. Ou seja, tem-se $Q(x)$.
   - $x = awa$ (ou $x = bwb$) para algum $w \in A^*$. Dado que $x \in L_1$, tem-se $|x|_a = |x|_b$. Logo $x = aw_1w_2a$, com $aw_1, w_2a \in L_1$ (porquê?). Pela H.I., tem-se $aw_1, w_2a \in L$. Assim, pela regra 4 da definição de L, $x \in L$. Ou seja, tem-se $Q(x)$.

De 1 e 2 segue $Q(x)$ para cada $x \in L_1$, i.e., $L_1 \subseteq L$.

Definição

Dado que linguagens são conjuntos de palavras, ficam imediatamente definidas para linguagens as diversas operações sobre conjuntos. Em particular, dadas linguagens L e K sobre A:

1. $L \cup K = \{u \in A^* : u \in L$ ou $u \in K\}$ - união de L e K.
2. $L \cap K = \{u \in A^* : u \in L$ e $u \in K\}$ - interseção de L e K.
3. $K \setminus L = \{u \in A^* : u \in K$ e $u \notin L\}$ - complementar de L em K.
4. $\overline{L} = A^* \setminus L = \{u \in A^* : u \notin L\}$ - complementar de L.

E.g., para $A = \{0,1\}$, $L = \{u \in A^* : |u|_0 = 0\}$ e $K = \{u \in A^* : |u|_1 = 0\}$:

$L \cup K = A^*$; $L \cap K = \emptyset$; $L \setminus K = L$; $\overline{L} = K$.

Observação

As anteriores operações em linguagens herdam, de imediato, as propriedades das respetivas operações em conjuntos. Por exemplo: a união e a interseção de linguagens são operações associativas, comutativas, que possuem elemento neutro e elemento absorvente.

Definição

Dadas linguagens L e K sobre A, a concatenação de L e K é a linguagem em A, notada por $L.K$ ou LK, dada por:

$$\{u.v : u \in L \text{ e } v \in K\}\,.$$

Por exemplo, para $L = \{b, ba\}$ e $K = \{\epsilon, a\}$:

$$\begin{aligned} L.K &= \{b, ba, baa\} \\ K.L &= \{b, ba, ab, aba\} \end{aligned}$$

Deste exemplo, pode deduzir-se que a concatenação de linguagens não é comutativa.

Proposição

A concatenação de linguagens é uma operação associativa, com elemento neutro $\{\epsilon\}$ e com a linguagem vazia como elemento absorvente, que distribui pela união, i.e., para $L, K, M \subseteq A^*$:

$$L.(K \cup M) = (L.K) \cup (L.M) \qquad (K \cup M).L = (K.L) \cup (M.L)$$

Observação

Adiante, uma palavra u será utilizada muitas vezes como uma abreviatura para $\{u\}$ (a linguagem cuja única palavra é u).

Por exemplo, considerando o alfabeto $A = \{a, b\}$,

$$ab\{aa, bb\} = \{ab\}\{aa, bb\} = \{abaa, abbb\} .$$

Proposição

Dada uma palavra u sobre um alfabeto A:

$$\begin{array}{rcl} uA^* & = & \{ux : x \in A^*\}, \\ A^*u & = & \{xu : x \in A^*\}, \\ A^*uA^* & = & \{xuy : x, y \in A^*\} \end{array}$$

são as linguagens cujas palavras têm, respetivamente, u como prefixo, como sufixo e como fator.

Definição

Dada uma linguagem L, a sua linguagem inversa será notada por L^I, sendo dada por:

$$L^I = \{u^I : u \in L\}.$$

Por exemplo, considerando o alfabeto $\{0, 1\}$:

1. para $L = \{\epsilon, 0, 01\}$, $L^I = \{\epsilon, 0, 10\}$;
2. para $K = \{0^n 1^m : n \in \mathbb{N}_0 \wedge m \in \mathbb{N}\}$, $K^I = \{1^m 0^n : n \in \mathbb{N}_0 \wedge m \in \mathbb{N}\}$.

Proposição

Dadas linguagens L e K:

1. $(L \cup K)^I = L^I \cup K^I$;
2. $(L.K)^I = K^I.L^I$;

Definição

Sejam L uma linguagem num alfabeto A e $n \in \mathbb{N}_0$. A potência n de L é a linguagem em A, notada por L^n, dada recursivamente por:

$$\begin{aligned} L^0 &= \{\epsilon\} \\ L^{k+1} &= L^k.L \quad (k \in \mathbb{N}_0) \end{aligned}$$

Por exemplo, para a linguagem $L = \{a, ab\}$ (sobre o alfabeto $A = \{a, b\}$):

$$\begin{aligned} L^0 &= \{\epsilon\} \\ L^1 &= L^0.L = \{\epsilon\}.\{a, ab\} = \{a, ab\} \\ L^2 &= L^1.L = \{a, ab\}.\{a, ab\} = \{aa, aab, aba, abab\} \quad . \end{aligned}$$

Proposição

Sejam u uma palavra, L uma linguagem num alfabeto A e $n \in \mathbb{N}_0$. Então, $u \in L^n$ se e só se

1. (i) $n = 0$ e $u = \epsilon$; ou
2. (ii) $n \geq 1$ e existem palavras $u_1, ..., u_n \in L$ t.q. $u = u_1 ... u_n$.

Definição

Dada uma linguagem L num alfabeto A:

1. o fecho positivo de L é a linguagem em A, notada por L^+, dada por:

$$L^+ = \bigcup_{n \geq 1} L^n$$

2. o fecho (de Kleene) de L, também designada a estrela de L, é a linguagem em A, notada por L^*, dada por:

$$L^* = \bigcup_{n \geq 0} L^n = L^+ \cup \{\epsilon\}.$$

Por exemplo, para o alfabeto $A = \{0, 1\}$ e para a linguagem $L = \{0, 1\}$ sobre A:

1. para todo $n \in \mathbb{N}_0$, $L^n = \{u \in A^* : |u| = n\}$(exercício);
2. $L^+ = A^+$ (porquê?);
3. $L^* = A^*$ (porquê?).

De facto, os três itens anteriores são válidos para qualquer alfabeto A.

Proposição

Seja L uma linguagem. Então,

1. $\emptyset^* = \{\epsilon\}$, $\emptyset^+ = \emptyset$, $\{\epsilon\}^* = \{\epsilon\} = \{\epsilon\}^+$;
2. $L = L^1 \subseteq L^+ \subseteq L^+ \cup \{\epsilon\} = L^*$;
3. $\epsilon \in L^+$ se e só se $\epsilon \in L$;
4. $L^+ = LL^* = L^*L$.

Definição

O conjunto das expressões regulares sobre um alfabeto A é o conjunto $\mathrm{ER}(A)$, de palavras sobre o alfabeto $A \cup \{\emptyset, \epsilon, (,), +, \cdot, {}^*\}$, definido indutivamente por:

1. $\emptyset \in \mathrm{ER}(A)$ e $\epsilon \in \mathrm{ER}(A)$;
2. $a \in \mathrm{ER}(A)$, para cada $a \in A$;
3. Se $r, s \in \mathrm{ER}(A)$, então $(r + s) \in \mathrm{ER}(A)$;
4. Se $r, s \in \mathrm{ER}(A)$, então $(r \cdot s) \in \mathrm{ER}(A)$;
5. Se $r \in \mathrm{ER}(A)$, então $(r^*) \in \mathrm{ER}(A)$.

Habitualmente, notaremos expressões regulares por: $r, s, r', r_0, \ldots$

Por exemplo, sendo $A = \{a, b\}$, são expressões regulares sobre A:

$$\emptyset,\ \epsilon,\ a,\ b,\ (a+b),\ ((b\cdot\emptyset)+\epsilon),\ ((a+b)^*),\ ((a^*)+(b^*)).$$

Observação

A notação das expressões regulares pode ser abreviada da seguinte forma:

- o símbolo $\cdot$ pode ser omitido;
- podem omitir-se parênteses desnecessários usando associatividade para as operações $+$ e $\cdot$, e considerando que * tem a maior prioridade e que $\cdot$ tem prioridade em relação a $+$;
- para $r \in ER(A)$ e $n \in \mathbb{N}_0$, a abreviatura r^n é definida recursivamente, por:
  - $r^0 = \epsilon, \quad r^1 = r$ e, para $n \geq 2$, $r^n = (r^{n-1} \cdot r)$;
  - $r^+ = (r \cdot (r^*))$.

Por exemplo,

- a^+b^2 é uma abreviatura de $((a \cdot (a^*)) \cdot (b \cdot b))$;
- $\emptyset^* a + (b + \epsilon)b$ abrevia $(((\emptyset^*) \cdot a) + (((b + \epsilon) \cdot b)))$.

Definição

A cada expressão regular r sobre um alfabeto A associa-se uma linguagem $\mathcal{L}(r)$ sobre A, dita a linguagem representada por r ou a linguagem de r. A função

$$\begin{array}{rrcl} \mathcal{L}: & ER(A) & \rightarrow & \mathcal{P}(A^*) \\ & r & \mapsto & \mathcal{L}(r) \end{array}$$

é definida recursivamente por:

1. $\mathcal{L}(\emptyset) = \emptyset$ e $\mathcal{L}(\epsilon) = \{\epsilon\}$;
2. $\mathcal{L}(a) = \{a\}$, para cada $a \in A$;
3. $\mathcal{L}((r+s)) = \mathcal{L}(r) \cup \mathcal{L}(s)$, para quaisquer $r, s \in ER(A)$;
4. $\mathcal{L}((r \cdot s)) = \mathcal{L}(r) \cdot \mathcal{L}(s)$, para quaisquer $r, s \in ER(A)$;
5. $\mathcal{L}((r)^*) = \mathcal{L}(r)^*$, para cada $r \in ER(A)$.

Por exemplo, sendo $A = \{a, b\}$, tem-se:

1. $\mathcal{L}((b+\epsilon)a) = \mathcal{L}(b+\epsilon)\mathcal{L}(a) = (\mathcal{L}(b) \cup \mathcal{L}(\epsilon))\{a\} = (\{b\} \cup \{\epsilon\})\{a\} = \{b, \epsilon\}\{a\} = \{ba, a\}$;
2. $\mathcal{L}(a^*) = \mathcal{L}(a)^* = \{a\}^* = \{a^n : n \in \mathbb{N}_0\}$;
3. $\mathcal{L}(a^*(a^3 + b)) = \{a\}^*\{a^3, b\} = \{a^m : m \geq 3\} \cup \{a^n b : n \in \mathbb{N}_0\}$;
4. $\mathcal{L}(a^* + b^*) = \mathcal{L}(a)^* \cup \mathcal{L}(b)^* = \{a\}^* \cup \{b\}^* = \{a^n : n \in \mathbb{N}_0\} \cup \{b^n : n \in \mathbb{N}_0\}$;
5. $\mathcal{L}((a+b)^*) = \mathcal{L}(a+b)^* = (\{a\} \cup \{b\})^* = \{a, b\}^* = A^*$, ou seja, o conjunto de todas as palavras sobre o alfabeto A;
6. $\mathcal{L}((a+b)^*aba(a+b)^*) = A^*abaA^*$ é a linguagem das palavras que têm aba como fator.

Definição

1. Uma linguagem L sobre um alfabeto A diz-se regular quando pode ser representada por alguma expressão regular sobre A, ou seja:

$$\exists r \in ER(A).\, L = \mathcal{L}(r)\,.$$

2. O conjunto das linguagens regulares sobre um alfabeto A será denotado por $Reg(A)$.

Por exemplo, todas as linguagens do slide anterior são regulares (dado serem iguais a $\mathcal{L}(r)$, para cada uma das expressões regulares consideradas).

São também exemplos de linguagens regulares $\emptyset$ e $\{\epsilon\}$ já que

$$\emptyset = \mathcal{L}(\emptyset) \quad \text{e} \quad \{\epsilon\} = \mathcal{L}(\epsilon).$$

Observação

Alternativamente, o conjunto $Reg(A)$ das linguagens regulares sobre A pode ser definido indutivamente por:

1. $\emptyset, \{\epsilon\} \in Reg(A)$;
2. $\{a\} \in Reg(A)$, para todo o $a \in A$;
3. $Reg(A)$ é fechado para as operações de união, concatenação e fecho de Kleene, ou seja,

$$\text{se } L, K \in Reg(A) \text{ então } L \cup K,\ L \cdot K,\ L^* \in Reg(A).$$

Observação

Note-se que, se L é uma linguagem regular sobre um alfabeto A, então $L^+ = L^*L$ também é uma linguagem regular sobre A.

Observação

1. Para um alfabeto finito $A = \{a_1, ..., a_n\}$:
   - a linguagem A é regular: $A = \{a_1\} \cup ... \cup \{a_n\} = \mathcal{L}(a_1 + ... + a_n)$;
   - a linguagem A^* é regular: $A^* = \{a_1, ..., a_n\}^* = \mathcal{L}((a_1 + ... + a_n)^*)$
2. Para qualquer palavra u sobre um alfabeto, $\{u\}$ é uma linguagem regular sobre A. De facto:
   - caso $u = \epsilon$: $\{u\} = \{\epsilon\} = \mathcal{L}(\epsilon)$;
   - caso $u = a_1 a_2 \cdots a_n$, com $a_1, a_2, \ldots, a_n \in A$:

   $$\{u\} = \{a_1\}\{a_2\}\cdots\{a_n\} = \mathcal{L}(a_1 a_2 \ldots a_n) = \mathcal{L}(u)\,.$$
3. Toda a linguagem finita L, sobre um alfabeto A, é regular. De facto:
   - caso $L = \emptyset$: $L = \emptyset = \mathcal{L}(\emptyset)$;
   - caso $L = \{u_1, \ldots, u_k\}$, com $k \geq 1$, $u_i \in A^*$:

   $$L = \{u_1\} \cup \cdots \cup \{u_k\} = \mathcal{L}(u_1 + \ldots + u_k).$$

Proposição

Existem linguagens que não são regulares.

A proposição segue por razões de cardinalidade. De facto:

- por um lado, o conjunto $\mathcal{P}(A^*)$ das linguagens sobre um alfabeto A finito, com duas ou mais letras, é infinito não numerável;
- por outro, o conjunto $ER(A)$ das expressões regulares sobre um alfabeto A finito é numerável.

Assim, há linguagens que não poderão corresponder à representação de qualquer expressão regular.

Observação

Prova-se que não são regulares as linguagens (sobre $A = \{0, 1\}$):

1. $\{0^p : p > 0 \text{ primo}\}$
2. $\{0^n 1^n : n \in \mathbb{N}_0\}$
3. $\{u \in A^* : u^I = u\}$

Note-se que expressões regulares distintas podem representar a mesma linguagem.

Por exemplo, as expressões regulares

$$(a+b)^* \quad \text{e} \quad (a+b)(a+b)^* + \epsilon$$

representam a mesma linguagem, nomeadamente $\{a, b\}^*$.

Definição

- Diremos que r é menor ou igual que s, escrevendo, $r \leq s$, quando $\mathcal{L}(r) \subseteq \mathcal{L}(s)$.
- Diremos que r é equivalente a s ou, simplesmente, que r é igual a s, escrevendo, respetivamente, $r \equiv s$ e $r = s$, quando $r \leq s$ e $s \leq r$, ou seja, quando $\mathcal{L}(r) = \mathcal{L}(s)$.

Observação

Dado um alfabeto A, $(ER(A), \leq)$ constitui um conjunto parcialmente ordenado. (Porquê?)

Por exemplo, considerando o alfabeto $A = \{a, b\}$, pode escrever-se:

1. $a \leq a + b$, mas $a + b \not\leq a$, pelo que $a \neq a + b$;
2. $a + b \leq (a + b)^+ \leq (a + b)^*$ e consequentemente $a + b \leq (a + b)^*$ (porquê?);
3. $(a + b)^* aa(a + b)^* \leq (a + b)^* a(a + b)^*$ (porquê?);
4. $(a + b)^* \leq (a + b)(a + b)^* + \epsilon$;
5. $(a + b)(a + b)^* + \epsilon \leq (a + b)^*$;
6. $(a + b)^* = (a + b)(a + b)^* + \epsilon$;

Observação

Adiante, por norma, dadas expressões regulares $r, s \in ER(A)$, a notação $r = s$ significará que r é equivalente a s, ou seja, $\mathcal{L}(r) = \mathcal{L}(s)$ (e não que as palavras r e s, sobre o alfabeto $A \cup \{\emptyset, \epsilon, (,), +, \cdot, {}^*\}$, são sequências de letras de iguais).

Proposição

Sejam r, s e t expressões regulares sobre um alfabeto A. Então,

(i) $(r+s)+t = r+(s+t)$;

(ii) $r+\emptyset = \emptyset + r = r$;

(iii) $r+s = s+r$;

(iv) $r+r = r$;

(v) $r\emptyset = \emptyset r = \emptyset$;

(vi) $r\epsilon = \epsilon r = r$;

(vii) $(rs)t = r(st)$;

(viii) $r(s+t) = rs+rt$;

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

(x) $\emptyset^* = \epsilon^* = \epsilon$;

(xi) $r^* = r^*r^* = (r^*)^* = (\epsilon+r)^* = r^+ + \epsilon$;

(xii) $r^+ = rr^* = r^*r$;

(xiii) $(r+s)^* = (r^*+s^*)^* = (r^*s^*)^* = (r^*s)^*r^*$;

(xiv) $r(sr)^* = (rs)^*r$;

(xv) $(r^*s)^* = (r+s)^*s+\emptyset$;

(xvi) $(rs^*)^* = r(r+s)^*+\epsilon$.

Definição

Uma equação linear à direita sobre expressões regulares é uma equação do tipo

$$X = rX + s$$

na qual $r, s \in ER(A)$ são expressões regulares e X é dita a (expressão regular) indeterminada ou incógnita.

Habitualmente, usaremos X, Y, X_1, ... para representar expressões regulares indeterminadas

Exemplos, de tais equações lineares são:

$$\begin{aligned} X &= aX + a \\ Y &= (a + b^*)Y + (a + \epsilon) \end{aligned}$$

Definição

Diz-se que uma expressão regular $t \in ER(A)$ é uma solução da equação $X = rX + s$ quando $t = rt + s$.

Por exemplo, uma solução da equação

$$X = aX + a$$

é a^+, dado que se tem $a^+ = aa^+ + a$, uma vez que:

$$\mathcal{L}(a^+) = \{a^n : n \in \mathbb{N}\} = \{a^n : n \geq 2\} \cup \{a\} = \mathcal{L}(aa^+) \cup \mathcal{L}(a) = \mathcal{L}(aa^+ + a).$$

$(a + b)^*$ é uma solução da equação $Y = (a + b^*)Y + (a + \epsilon)$.
(Porquê?)

É possível que uma equação linear à direita tenha várias soluções. Por exemplo, a equação

$$X = \epsilon X + a + b$$

tem como solução $a + b$ (pois: $\mathcal{L}(a+b) = \{a, b\} = \mathcal{L}(\epsilon(a+b) + a+b)$), bem como $a+b+\epsilon$ (pois: $\mathcal{L}(a+b+\epsilon) = \{a,b,\epsilon\} = \mathcal{L}(\epsilon(a+b+\epsilon)+a+b)$). De facto, toda a expressão regular r tal que $\mathcal{L}(r) \supseteq \{a, b\}$ é solução desta equação (porquê?). Na verdade, esta condição é necessária para r ser solução desta equação (porquê?) e, por esta razão, $a + b$ dir-se-á solução mínima desta equação:

Definição

Diz-se que uma expressão regular $t \in ER(A)$ é uma solução mínima da equação $X = rX + s$ quando:

1. t é uma solução da equação ; e,
2. para toda a solução t' desta equação, $t \leq t'$.

Proposição

Sejam $r, s \in ER(A)$.

(a) Se $t, t' \in ER(A)$ são soluções mínimas da equação $X = rX + s$, então $t = t'$.

(b) $r^* s$ é a solução mínima da equação $X = rX + s$.

(c) Se $\epsilon \notin \mathcal{L}(r)$, então $r^* s$ é a única solução de $X = rX + s$.

Por exemplo, a solução mínima da equação

$$X = (a + b)X + \epsilon$$

é $(a + b)^*$, por (b). Dado que $\epsilon \notin \mathcal{L}(a + b)$, então, por (c), $(a + b)^*$ é a única solução desta equação.

Demonstração da Proposição: Provaremos apenas (a) e (b).

(a) Suponhamos que $t, t' \in ER(A)$ são soluções mínimas da equação $X = rX + s$. Então, $t \leq t'$, pois t é solução mínima, e $t' \leq t$, pois t' é solução mínima. Logo, $t = t'$.

(b) (i) Tem-se: $r(r^*s) + s = (rr^*)s + s = (rr^* + \epsilon)s = r^*s$. Portanto, r^*s é solução de $X = rX + s$.

(ii) Seja $t \in ER(A)$ outra solução de $X = rX + s$. Então, $t = rt + s$, o que significa que

$$\mathcal{L}(t) = \mathcal{L}(r)\mathcal{L}(t) \cup \mathcal{L}(s). \tag{1}$$

Logo, $\mathcal{L}(r)\mathcal{L}(t) \subseteq \mathcal{L}(t)$. Daqui decorre que

$$\mathcal{L}(r)^2\mathcal{L}(t) \subseteq \mathcal{L}(r)\mathcal{L}(t) \subseteq \mathcal{L}(t)$$

e, indutivamente, tem-se $\mathcal{L}(r)^n\mathcal{L}(t) \subseteq \mathcal{L}(t)$, para todo o $n \in \mathbb{N}_0$. Portanto, $\mathcal{L}(r)^*\mathcal{L}(t) = \mathcal{L}(t)$. Usando a igualdade (1), deduz-se:

$$\mathcal{L}(t) = \mathcal{L}(r)^*(\mathcal{L}(r)\mathcal{L}(t) \cup \mathcal{L}(s)) = \mathcal{L}(r)^*\mathcal{L}(r)\mathcal{L}(t) \cup \mathcal{L}(r)^*\mathcal{L}(s).$$

Conclui-se, então, que $\mathcal{L}(r)^*\mathcal{L}(s) \subseteq \mathcal{L}(t)$, donde $r^*s \leq t$. Portanto, r^*s é solução mínima da equação $X = rX + s$, que, por (a), é única.

Definição

Um sistema de equações lineares à direita sobre expressões regulares é um sistema da forma

$$\begin{cases} X_1 = r_{11}X_1 + r_{12}X_2 + \cdots + r_{1n}X_n + s_1 \\ X_2 = r_{21}X_1 + r_{22}X_2 + \cdots + r_{2n}X_n + s_2 \\ \quad\vdots \\ X_n = r_{n1}X_1 + r_{n2}X_2 + \cdots + r_{nn}X_n + s_n \end{cases}$$

onde $r_{ij}, s_i \in ER(A)$ para todos os $i, j \in \{1, \ldots, n\}$ e $X_1, X_2, \ldots, X_n$ são chamadas as indeterminadas ou incógnitas.

Diz-se que:

- $(t_1, t_2, \ldots, t_n) \in ER(A)^n$ é uma solução do sistema quando, para cada $i \in \{1, \ldots, n\}$, $t_i = r_{i1}t_1 + r_{i2}t_2 + \cdots + r_{in}t_n + s_i$;
- uma solução $(t_1, t_2, \ldots, t_n) \in ER(A)^n$ do sistema é uma solução mínima quando, para toda a solução $(t'_1, t'_2, \ldots, t'_n)$ do sistema, $t_i \leq t'_i$ para todo o $i \in \{1, \ldots, n\}$.

Proposição

(a) Um sistema de equações lineares à direita sobre expressões regulares, num dado alfabeto, tem uma única solução mínima.

(b) Se $\epsilon \notin \mathcal{L}(r_{ij})$, para cada coeficiente r_{ij} do sistema, então o sistema tem uma única solução.

Observação

Para determinar a solução mínima de um sistema pode usar-se:

- o “método de substituição” e
- a solução mínima das equações da forma $X = rX + s$.

Consideremos, por exemplo, o sistema

$$\begin{cases} X_1 = bX_1 + aX_2 + \emptyset \\ X_2 = aX_1 + bX_2 + \epsilon \end{cases}$$

e determinemos a sua solução mínima. Pode deduzir-se, sucessivamente:

$$\begin{aligned}
\begin{cases} X_1 = bX_1 + aX_2 + \emptyset \\ X_2 = aX_1 + bX_2 + \epsilon \end{cases}
&\Leftrightarrow \begin{cases} X_1 = b^* aX_2 \\ X_2 = aX_1 + bX_2 + \epsilon \end{cases} \\
&\Leftrightarrow \begin{cases} X_1 = b^* aX_2 \\ X_2 = ab^* aX_2 + bX_2 + \epsilon \end{cases} \\
&\Leftrightarrow \begin{cases} X_1 = b^* aX_2 \\ X_2 = (ab^* a + b)X_2 + \epsilon \end{cases} \\
&\Leftrightarrow \begin{cases} X_1 = b^* a(ab^* a + b)^* \\ X_2 = (ab^* a + b)^* \end{cases}
\end{aligned}$$

A solução mínima do sistema é, portanto:

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

Observação

Os sistemas de equações lineares podem ser usados para determinar uma expressão regular que represente uma dada linguagem (regular).

Por exemplo, para $A = \{a, b\}$, sendo

$$L_1 = \{u \in A^* : \ |u|_a \text{ é ímpar}\} \quad \text{e} \quad L_2 = \{u \in A^* : \ |u|_a \text{ é par}\}$$

são válidas as igualdades $L_1 = bL_1 \cup aL_2$ e $L_2 = aL_1 \cup bL_2 \cup \{\epsilon\}$.
Ou seja, (L_1, L_2) é a única solução do sistema

$$\begin{cases} X_1 = bX_1 \cup aX_2 \cup \emptyset \\ X_2 = aX_1 \cup bX_2 \cup \{\epsilon\} \end{cases}$$

que convertido em sistema de equações lineares é precisamente o sistema do exemplo anterior. Portanto, considerando a solução mínima desse sistema, deduz-se que

$$r_1 = b^*a(ab^*a + b)^* \quad \text{e} \quad r_2 = (ab^*a + b)^*$$

são expressões regulares que representam L_1 e L_2, respetivamente.