IUT de Bordeaux
Département Informatique

M2201 – Graphes et Langages

Éléments de Théorie des Langages

- **Introduction générale : alphabets, mots et langages**
- **Langages rationnels**

Alphabet, mot sur un alphabet (1)

Un **alphabet** est un ensemble (fini) **A** de symboles appelés **lettres**.

Un **mot** m sur un alphabet **A** est une séquence (finie) de lettres prises dans **A** : $\mathbf{m = a_1 \ldots a_k}$. Ce mot est de **longueur** (nombre de lettres) k : $\mathbf{|m| = k}$.

Il existe un unique mot de longueur nulle, le **mot vide**, noté ε.

On note $\mathbf{A}^*$ l'ensemble de tous les mots construits sur l'alphabet **A**. On note $\mathbf{A}^n$ l'ensemble des mots de $\mathbf{A}^*$ de longueur n.

En particulier, $\mathbf{A}^0 = \{ \varepsilon \}$ et $\mathbf{A}^1 = \mathbf{A}$.

Alphabet, mot sur un alphabet (2)

Soit a une lettre de **A** et m un mot de **A***. Le **nombre d'occurrences** de la lettre a dans m, noté $|m|_a$, est le nombre de fois où la lettre a apparaît dans m.

Notons que $|\varepsilon| = 0$ et, pour toute lettre a, $|\varepsilon|_a = 0$.

Exemples.

$\mathbf{A_1} = \{ 0, 1 \}$

$m_1 = 00101011, \quad m_2 = 1101$

$|m_1| = 8, \quad |m_2|_1 = 3$

Alphabet, mot sur un alphabet (3)

$\mathbf{A_2}$ = { a, b, c }

m_3 = baba, m_4 = bac

A_3 = { 0, …, 9, +, -, *, :, (,) }

m_5 = (12 + 4) * (71 – 14:5)

A_4 = { si, alors, sinon, >, a, b, ←, +, 0, 1 , … }

m_6 = si a > b + 1 alors a ← 0 sinon b ← 10

Concaténation de mots

Soient u et v deux mots de **A***. La **concaténation** de u et v est le mot, noté u.v ou plus simplement uv, obtenu en « collant » le mot v à la suite du mot u. Ainsi, $|uv| = |u| + |v|$ et, pour toute lettre a, $|uv|_a = |u|_a + |v|_a$.

On notera u^n le mot **u.u. … .u** (n fois), avec $u^0 = \varepsilon$.

Exemple. u = aba, v = ca, uv = abaca, u^2 = abaaba

$|u| = 3$, $|v| = 2$, $|uv| = 5$, $|u^2| = 6$

Pour tous mots u, v et w, nous avons :

- **εu = uε = u** (ε est élément neutre)
- **u.vw = uv.w = uvw** (associativité)
- mais, en général, **uv ≠ vu** (non commutativité)

Préfixes, suffixes et facteurs

Soient u et v deux mots de **A***.

Le mot u est un **préfixe** du mot v s'il existe un mot w de **A*** tel que v = uw.

De façon similaire, le mot w est un **suffixe** du mot v s'il existe un mot u de **A*** tel que v = uw.

Le mot u est un **facteur** du mot v s'il existe deux mots w_1 et w_2 de **A*** tels que $v = w_1uw_2$.

Exemple. u = aba, v = abac, w = abacabac, t = aca

- u est un *préfixe* de w, car w = u.cabac
- v est un *suffixe* de w, car w = abac.v
- t est un *facteur* de w, cat w = ab.t.bac

Quelques propriétés…

Propriété. Si u, v et w sont trois mots de A*, alors uw = vw $\Leftrightarrow$ wu = wv $\Leftrightarrow$ u = v

Lemme de Levi. Si u et v sont tous deux préfixes de w, alors u est préfixe de v, ou v est préfixe de u.

Théorème (de commutation). Si u et v commutent (c'est-à-dire sont tels que uv = vu), alors u et v sont deux *puissances* d'un même facteur :

$\Rightarrow$ il existe un mot f de **A*** et deux entiers p et q tels que $u = f^p$ et $v = f^q$.

Preuve du théorème de commutation…

- Si $u = f^p$ et $v = f^q$, alors $uv = vu = f^{p+q}$.
- Réciproque : par récurrence sur $N = |u| + |v|$:
  - si $N = 0$, alors $u = v = \varepsilon$ et donc $uv = vu$.
  - si $u = \varepsilon$, alors $u = v^0$ et $v = v^1$ (idem si $v = \varepsilon$).
  - si $|u| = |v|$, alors $uv = vu$ entraîne $u = v$
    d'où $f = u$, $p = q = 1$.
  - sinon, comme u et v sont préfixes de $uv = vu$, l'un est préfixe de l'autre (Lemme de Levi).

    Supposons que u est préfixe de v : $v = uw$.

    On a donc $uv = vu \Rightarrow u(uw) = (uw)u$, c'est-à-dire
    $uuw = uwu$, et donc $uw = wu$ (propriété).

    L'hypothèse de récurrence permet de conclure.

Langages (1)

Un **langage** L sur un alphabet **A** est un sous-ensemble, fini ou infini, de **A***.

Par exemple, sur l'alphabet **A** = {a, b}, on peut définir les langages suivants :

- $L_1 = \mathbf{A}^2 = \{ aa, ab, ba, bb \}$,
- L_2 = { mots d'au plus quatre lettres ayant autant de a que de b }

  = {ε, ab, ba, aabb, abab, abba, baab, baba, bbaa},
- L_3 = { mots ayant deux fois plus de a que de b } (ce langage est infini).

Langages (2)

Autres exemples :

- A = { a, …, z }

  L = { mots de la langue française }

- A = { mots de la langue française }

  L = { phrases correctes }

- A = { … }

  L = { programmes C++ syntaxiquement corrects }

- etc.

Opérations sur les langages (1)

Soit **A** un alphabet. On définit les opérations suivantes sur les langages définis sur **A*** :

Union.

$$L_1 \cup L_2 = \{ m \in \mathbf{A}^* / (m \in L_1) \text{ ou } (m \in L_2) \}$$

Intersection.

$$L_1 \cap L_2 = \{ m \in \mathbf{A}^* / (m \in L_1) \text{ et } (m \in L_2) \}$$

Produit (de concaténation).

$$L_1.L_2 = L_1L_2 = \{m \in \mathbf{A}^* / m = m_1m_2,\ m_1 \in L_1 \text{ et } m_2 \in L_2\}$$

Opérations sur les langages (2)

Puissance.

$$L^0 = \{ \varepsilon \}$$

$$L^n = \{ m \in \mathbf{A}^* / m = m_1 m_2 ... m_n,$$

$$\text{avec } m_i \in L \text{ pour tout } i,\ 1 \leq i \leq n \}, \text{ pour } n \geq 1$$

Étoile et "plus".

$$L^* = L^0 \cup L^1 \cup ... \cup L^n \cup ...$$

$$L^+ = L^1 \cup L^2 \cup ... \cup L^n \cup ...$$

$$(\text{i.e. } L^+ = L.L^*)$$

Langages rationnels

Langages rationnels (1)

Un langage sur un alphabet **A** est **rationnel** (on dit également **régulier**), s'il peut être construit à l'aide des opérations union, étoile et produit à partir des langages élémentaires composés du mot vide ou d'un mot à une seule lettre.

De façon inductive, un L.R. se définit donc ainsi :

- $\{ \varepsilon \}$ est un L.R.,
- si $a \in \mathbf{A}$, $\{ a \}$ est un L.R.,
- si L est un L.R. alors L^* est un L.R. (L^+ et L^n aussi),
- si L_1 et L_2 sont des L.R., alors $L_1 \cup L_2$ et $L_1.L_2$ sont des L.R.

Langages rationnels (2)

Soit **A** = { a, b, c }. Les langages suivants sont rationnels :

- $L_1 = (\{ \varepsilon \} \cup \{ a \}).(\{ b \} \cup \{ c \})^*.\{ a \}^3$

  $L_1 = \{ aaa, bcbbcaaa, ..., aaaa, abaaa, ... \}$

- $L_2 = \{ c \}^2.(\{ a \} \cup \{ b \}^* \cup (\{ a \} \cup \{ b \}.\{ c \})^2)^3$

  $L_2 = \{ ccaaa,cc,cca,ccaaaaaa,ccbcabcabcbc, ... \}$

- $L_3 = (\{ a \} \cup \{ b \})^*.\{ c \}$

  $L_3 = \{ c, ac, bc, aac, abc, bac, bbc, aaac, ... \}$

Expressions rationnelles

On utilise habituellement quelques conventions d'écriture qui permettent d'alléger les expressions précédentes :

- un langage singleton **{ a }** est noté simplement **a**,
- l'union est notée simplement **+**.

Ainsi, les langages précédents peuvent s'écrire, plus simplement, sous forme de ce que l'on appelle des **expressions rationnelles** :

- $L_1 = (\varepsilon + a)(b + c)^* a^3$
- $L_2 = c^2(a + b^* + (a + bc)^2)^3$
- $L_3 = (a + b)^* c$