X5I0050 - Langages et automates
Généralités sur les langages

D. Béchet & T. Sadiki

Université de Nantes & Université Internationale de Rabat

10 septembre 2014

Introduction

Langage naturel, langage mathématique, langage de programmation (C, Java, ...), langage formel, format de fichiers, normalisation, ...

Langage formel : un vocabulaire + des règles de grammaires

Théorie des langages formels

- **Analyse** : déterminer si une phrase appartient ou non au langage.
- **Génération** : générer l'ensemble des phrases (mots) d'un langage

Syntaxe uniquement
Sémantique non prise en compte

Voir Noam CHOMSKY (sur Wikipédia) pour l'aspect historique

Plan Chapitre 1

1. Notions de monoïde
2. Mots et monoïdes
3. Relations entre mots
4. Langages et opérations

Définition de monoïde

Définition 1.1 - Monoïde

Un **monoïde** est un ensemble E muni d'une opération binaire interne associative $\oplus$ et possédant un élément neutre ε
Notation : $< E, \oplus, \varepsilon >$

Exemples :
- $< \mathbb{N}, +, 0 >$ les entiers positifs ou nuls avec l'addition (élément neutre 0)
- $< \mathbb{Z}, +, 0 >$ les entiers relatifs avec l'addition (élément neutre 0)
- $< \mathbb{R}, \times, 1 >$ les nombres rééls avec la multiplication (élément neutre 1)
- $< \mathcal{P}(U), \cup, \emptyset >$ les sous-ensembles d'un univers U avec l'union
- etc

Propriétés d'un monoïde $< E, \oplus, \varepsilon >$

- $\oplus$ est une **opération binaire interne** (ou **stable**) à E :

$$\forall x, y \in E, x \oplus y \text{ est défini } \text{ et } \; x \oplus y \in E$$

- $\oplus$ est **associative** :

$$\forall x, y, z \in E, (x \oplus y) \oplus z = x \oplus (y \oplus z)$$

  Pour cette raison, les parenthèses sont souvent omises

- ε est un **élément neutre** de E pour $\oplus$:

$$\varepsilon \in E \;\text{ et }\; \forall x \in E, \varepsilon \oplus x = x \oplus \varepsilon = x$$

- Il ne peut exister qu'un seul élément neutre dans un monoïde : si ε et ε' sont deux éléments neutres de E pour $\oplus$, alors $\varepsilon = \varepsilon \oplus \varepsilon' = \varepsilon'$

Définition de sous-monoïde

Définition 1.2 - Sous-monoïde

Soit $< E, \oplus, \varepsilon >$, un monoïde et T un sous-ensemble de E (T $\subseteq$ E).
$< T, \oplus, \varepsilon >$ est un **sous-monoïde** de $< E, \oplus, \varepsilon >$, ssi c'est un monoïde.
Puisque $< E, \oplus, \varepsilon >$ est un monoïde, il suffit de démontrer que :

- $\varepsilon \in$ T
- $\oplus$ est stable dans T : $\forall$ t,t' $\in$ T, t $\oplus$ t' $\in$ T

Définition 1.3 - Monoïde engendré

Soit M $= < E, \oplus, \varepsilon >$ un monoïde. Pour toute partie A de E, on peut définir le plus petit sous-monoïde de M contenant A. On l'appelle le **sous-monoïde de M engendré par A**.

Exercices sue les monoïdes

Monoïde ? Sous-monoïde ? Monoïde engendré ?

$\mathbb{N}$ est l'ensemble des entiers naturels
Pair($\mathbb{N}$) est l'ensemble des entiers pairs (ou nuls)
Impair($\mathbb{N}$) est l'ensemble des entiers impairs

- $< \mathbb{N}, +, 0 >$?
- $< \mathbb{N}, \times, 1 >$?
- $< Pair(\mathbb{N}), +, 0 >$?
- $< Pair(\mathbb{N}), \times, 1 >$?
- $< Impair(\mathbb{N}), +, 0 >$?
- $< Impair(\mathbb{N}), \times, 1 >$?

Exercices sue les monoïdes - Quelques réponses

Monoïde ? Sous-monoïde ? Monoïde engendré ?

1. $< \mathit{Pair}(\mathbb{N}), +, 0 >$: un monoïde
2. $< \mathit{Pair}(\mathbb{N}), +, 0 >$: sous-monoïde de $< \mathbb{N}, +, 0 >$
3. $< \mathit{Pair}(\mathbb{N}), +, 0 >$: le sous-monoïde de $< \mathbb{N}, +, 0 >$ engendré par $\{2\}$
4. $< \mathit{Impair}(\mathbb{N}), +, 0 >$: n'est pas un monoïde

Démonstrations :

- (3) implique (2) qui implique (1)
- (3) se démontre en remarquant que 2 est pair, en montrant (2) puis en montrant par une récurrence simple que tout sous-monoïde de $< \mathbb{N}, +, 0 >$ contenant l'entier 2 doit contenir au moins tous les entiers pairs
- (4) se démontre facilement en remaquant que la somme de deux nombres impairs n'est pas impair, par exemple $1 + 1 = 2$

Définitions de symbole, alphabet, mot

Définition 1.4 - Symbole

Un **symbole** est une brique élémentaire, un atome

Définition 1.5 - Alphabet

Un **alphabet** A est un ensemble fini et non vide de symboles

Définition 1.6 - Mot

Un **mot** (**chaîne**) sur un alphabet A est une suite finie de symboles de A
Si $a_1, \ldots, a_n$ sont des symboles de A, on note $\mathbf{a_1 \cdots a_n}$ le mot qui forme la suite de ces n symboles[a]

[a] n peut être nul et les symboles ne sont pas forcément distincts deux à deux

Longueur d'un mot, mot vide

Définition 1.7 - Longueur d'un mot

La **longueur** d'un mot est le nombre de symboles du mot. Soit un mot w, sa longueur est notée $|\mathbf{w}|$

Définition 1.8 - Mot vide

Le **mot vide** est le mot de longueur 0, c'est-à-dire ne contenant aucun symbole. Plutôt que de le marquer par une chaîne vide, il est souvent noté ε. $|\varepsilon| = 0$

Définition 1.9 - Mot de longueur n, A^n, A^* et A^+

Soit A un alphabet, l'ensemble des **mots de longueur n** est noté $\mathbf{A^n}$
On appelle $\mathbf{A}^*$, l'ensemble de tous les mots de longueur finie construits avec les symboles de A.
On appelle $\mathbf{A}^+$, l'ensemble de tous les mots de longueur finie et non nulle construits avec les symboles de A

Concaténation

Définition 1.10 - Concaténation

Soit $v, w \in A^*$, $|v| = m$, $|w| = n$, la **concaténation** de v et w, notée $\mathbf{v} \cdot \mathbf{w}$, est le mot de longueur $m + n$ dont les m premiers symboles sont le mot v et les n suivants le mot w. Ainsi, si $v = a_1 a_2 \ldots a_m$ et $v = b_1 b_2 \ldots b_n$ alors $\mathbf{v} \cdot \mathbf{w} = \mathbf{a_1 a_2 \ldots a_m b_1 b_2 \ldots b_n}$

Propriétés de la concaténation

- associativité, $\forall u, v, w \in A^*$, $(u \cdot v) \cdot w = u \cdot (v \cdot w) = u \cdot v \cdot w$
- généralement non commutatif, soit $u, v \in A^*$, $u \cdot v \neq v \cdot u$
- élément neutre ε, $\forall u \in A^*$, $\varepsilon \cdot u = u \cdot \varepsilon = u$

$< A^*, \cdot, \varepsilon >$ est un monoïde, appelé **monoïde libre engendré par A**

Décomposition

Théorème 1.1 - Propriété de la décomposition

Tout mot v (de longueur $n = |v|$) sur un alphabet A se décompose de façon unique en $a_1 \cdot a_2 \cdot \ldots \cdot a_n$, où $\forall i \in \{1 \ldots n\}, a_i \in A$.

Cela signifie que le monoïde libre engendré par A est un monoïde libre au sens de l'algèbre universelle, que les symboles sont atomiques et que tout mot est produit unique de concaténation des symboles qui le composent

Théorème 1.2 - Lemme de Levi

Soient $t, u, v, w \in A^*$, si $t \cdot u = v \cdot w$ alors il existe un unique $(\exists!)$ $z \in A^*$ tel que :

1. $(v = t \cdot z$ et $u = z \cdot w)$
2. ou bien $(t = v \cdot z$ et $z \cdot u = w)$

Ce lemme est un résultat intermédiaire utile dans les démonstrations

Démonstration du lemme de Levi

De manière informelle, si l'on a $t \cdot u = v \cdot w$, il y a trois cas :

1. $|t| = |v|$ et $|u| = |w|$:
   ```
   ttttttuuuuuu
         |
   vvvvvvwwwwww
   ```
   $\Longrightarrow z = \varepsilon$

2. $|t| < |v|$ et $|u| > |w|$:
   ```
   ttttuuuuuuuu
       |zzzz|
   vvvvvvvvwwww
   ```

3. $|t| > |v|$ et $|u| < |w|$:
   ```
   ttttttttuuuu
       |zzzz|
   vvvvwwwwwwww
   ```

Puissance

Notation - Puissance

La **puissance** n du mot x (notée $\mathbf{x^n}$) est le mot $v = x \cdot \ldots \cdot x$ (n fois).
On a $|v| = n \times |x|$

$\forall x \in A^*,\ x^0 = \varepsilon,\ x^1 = x$
$\forall x \in A^*,\ x^n = \varepsilon \Longrightarrow x = \varepsilon$ ou $n = 0$

Exercices sur les mots

1. Donner la longueur des mots suivants sur l'alphabet a, b, c :

   $abcb$
   $abba$
   ε
   bb

2. Donner la longueur des mots suivants sur l'alphabet $a, b, ch, ','$:

   $a,bbaa$
   $achbbaa$

3. Soient a, b des symboles de A et u un mot de A^*. Montrer que si $u \cdot a = b \cdot u$ alors $a = b$ et $u \in \{a\}^*$.

Exercices sur les mots - Réponses

1. Donner la longueur des mots suivants sur l'alphabet a,b,c :

   $$abcb \Longrightarrow 4$$
   $$abba \Longrightarrow 4$$
   $$\varepsilon \Longrightarrow 0$$
   $$bb \Longrightarrow 2$$

2. Donner la longueur des mots suivants sur l'alphabet $a,b,ch,','$:

   $$a,bbaa \Longrightarrow 6$$
   $$achbbaa \Longrightarrow 6$$

3. Soient a, b des symboles de A et u un mot de A^*. Montrer que si $u \cdot a = b \cdot u$ alors $a = b$ et $u \in \{a\}^*$
   Démonstration : par induction généralisée sur la longueur de u et en utilisant le lemme de Levi

Exercices sur les mots - Démonstration de (3)

- **Hypothèse d'induction** Soient $k \in \mathbb{N}, u \in A^*, a, b \in A$ avec $|u| = k$, nous supposons que la propriété est vraie pour tout mot u' de longueur plus petite (strictement) que k.
- **Hypothèse** : nous supposons que $u \cdot a = b \cdot u$
- **Utilisation du lemme de Levi** : nous avons soit $u = b \cdot z$ et $z \cdot a = u$ soit $b = u \cdot z$ et $z \cdot u = a$
- **Sous cas 1** : si $k = 0$ alors $u = \varepsilon$ donc $a = b$ et $u \in \{a\}^*$
- **Sous cas 2** : si $k > 0$, $b = u \cdot z$ et $z \cdot u = a$ alors $z = \varepsilon$ et $a = b = u$ donc $a = b$ et $u \in \{a\}^*$
- **Sous cas 3** : si $k > 0$, $u = b \cdot z$ et $z \cdot a = u$ alors $z \cdot a = b \cdot z$
  Utilisation de l'hypothèse d'induction : comme $z \cdot a = b \cdot z$ et $|z| < |u| = k$, nous avons $a = b$ et $z \in \{a\}^*$
  Finalement $a = b$ et $u = z \cdot a \in \{a\}^*$

Définitions de préfixe, suffixe, sous-chaîne/facteur

Définition 1.11 - Préfixe/préfixe propre

v est un **préfixe** de w ($\mathbf{v} \in \mathbf{Pref(w)}$) ssi $\exists z \in A^*$ tel que $w = v \cdot z$
v est un **préfixe propre** de w ($\mathbf{v} \in \mathbf{PrefProp(w)}$) ssi $\exists z \in A^+$ tel que $w = v \cdot z$

Définition 1.12 - Suffixe/suffixe propre

v est un **suffixe** de w ($\mathbf{v} \in \mathbf{Suff(w)}$) ssi $\exists z \in A^*$ tel que $w = z \cdot v$
v est un **suffixe propre** de w ($\mathbf{v} \in \mathbf{SuffProp(w)}$) ssi $\exists z \in A^+$ tel que $w = z \cdot v$

Définition 1.13 - Sous-chaîne

v est une **sous-chaîne** ou **facteur** de w ssi $\exists u_1, u_2 \in A^*$ tel que $w = u_1 \cdot v \cdot u_2$

facteur gauche $=$ préfixe
facteur droit $=$ suffixe

Occurrence et relations d'ordre sur les chaînes

Définition 1.14 - Occurrence

Une **occurrence** du symbole x dans le mot w est un entier $i > 0$ tel que le i-ème symbole de w (noté $w(i)$) est x
Le nombre d'occurrences de la lettre x dans le mot w est noté $|w|_x$

L'alphabet $A = \{x_1, x_2, \ldots, x_n\}$ muni d'un ordre total $<$ sur les symboles définit sur A^* plusieurs relations d'ordre :

- l'**ordre préfixiel**, noté $<_\mathbf{p}$: ordre partiel défini par : $\mathbf{u} <_\mathbf{p} \mathbf{v}$ ssi u est un préfixe propre de v
- l'**ordre lexicographique** (ordre du dictionnaire), noté $<_\mathbf{L}$: ordre total défini par : $\mathbf{u} <_\mathbf{L} \mathbf{v}$ ssi u est un préfixe propre de v ($u <_p v$) ou bien $u = w \cdot a \cdot u_2$ et $v = w \cdot b \cdot v_2$ tels que $w \in A^*$, $a < b$ avec $a, b \in A$
- l'ordre **hiérarchique**, noté $<_\mathbf{h}$: ordre total pour lequel les mots sont classés en premier lieu par longueur, puis pour les mots de même longueur, par ordre lexicographique.

Exercices sur les relations entre mots

1. Soit $x = abbcc$ un mot sur l'alphabet $A = \{a, b, c\}$, donner $Pref(x)$ et $Suff(x)$.
2. Soient $u_1, u_2, v \in A^*$, montrer que si $u_1 \in Pref(v)$ et $u_2 \in Pref(v)$ alors soit $u_1 \in Pref(u_2)$, soit $u_2 \in Pref(u_1)$
3. Soient les mots suivants sur l'alphabet $A = \{a, b, c, d\}$ muni de l'ordre total $a < b < c < d$:
    1. a
    2. abcd
    3. bc
    4. dbc
    5. ab
    6. cd
    7. cdab
    8. abdd

    Trier ces mots selon les ordres préfixiels, lexicographiques et hiérarchiques.

Exercices sur les relations entre mots - Réponses

1. Soit $x = abbcc$ un mot sur l'alphabet $A = \{a, b, c\}$
   $Pref(x) = \{a, ab, abb, abbc, abbcc\}$
   $Suff(x) = \{c, cc, bcc, bbcc, abbcc\}$
2. Soient $u_1, u_2, v \in A^*$, montrer que si $u_1 \in Pref(v)$ et $u_2 \in Pref(v)$ alors soit $u_1 \in Pref(u_2)$, soit $u_2 \in Pref(u_1)$

   Démonstration : assez simple en utilisant le lemme de Levi
   Si $u_1 \in Pref(v)$ et $u_2 \in Pref(v)$ alors $\exists w_1, w_2 \in A^*$ tels que $v = u_1 \cdot w_1 = u_2 \cdot w_2$. Le lemme de Levi implique que $\exists! z \in A^*$ tel que $(u_1 = u_2 \cdot z$ et $w_2 = z \cdot w_1)$ ou $(u_2 = u_1 \cdot z$ et $w_1 = z \cdot w_2)$
   $\Rightarrow u_1 = u_2 \cdot z$ ou $u_2 = u_1 \cdot z \Rightarrow u_1 \in Pref(u_2)$ ou $u_2 \in Pref(u_1)$

Exercices sur les relations entre mots - Réponses

3. Soient les mots suivants sur l'alphabet $A = \{a, b, c, d\}$ muni de l'ordre total $a < b < c < d$:
   - Ordre préfixiel :
     $a <_p abcd$, $a <_p ab$, $a <_p abdd$, $ab <_p abcd$, $ab <_p abdd$, $cd <_p cdab$
     $a <_p ab \begin{array}{l} <_p abcd \\ <_p abdd \end{array}$ et $cd <_p cdab$
   - Ordre lexicographiques :
     $a <_l ab <_l abcd <_l abdd <_l bc <_l cd <_l cdab <_l dbc$
   - Ordre hiérarchique :
     $a <_h ab <_h bc <_h cd <_h dbc <_h abcd <_h abdd <_h cdab$

Langage

Définition 1.15 - Langage

Un **langage** L sur un alphabet A est un ensemble de chaînes (ou ensemble de mots) sur A. L est donc un sous-ensemble de A^*, autrement dit $L \subseteq A^*$.

L'ensemble des langages L sur A est l'ensemble $\mathcal{P}(A^*)$ des parties de A^*, autrement dit : $L \in \mathcal{P}(A^*)$.
Étant donné un alphabet A, parmi tous les langages L de $\mathcal{P}(A^*)$:

- Le **langage neutre** est celui dont le seul mot est la chaîne vide : $\{\varepsilon\}$
- Le **langage vide** est celui qui ne contient aucun mot : $\emptyset$
- Un **langage fini** est un langage qui contient un nombre fini de mots
- Un **langage infini** est un langage qui n'est pas fini
- **Langage prefixe/suffixe** : un langage L est dit posséder la propriété préfixe (resp. suffixe) si aucune chaîne de L n'est préfixe (resp. suffixe) propre d'une autre chaîne de L

Opérateurs ensemblistes classiques et opérateurs induits par la concaténation des mots

Définition 1.16 - Union, intersection, différence, complémentaire

Union : $\mathbf{L} \cup \mathbf{M} = \{x \mid x \in L \text{ ou } x \in M\}$
Intersection : $\mathbf{L} \cap \mathbf{M} = \{x \mid x \in L \text{ et } x \in M\}$
Différence (ou exclusion) : $\mathbf{L} \backslash \mathbf{M} = L - M = \{x \mid x \in L \text{ et } x \notin M\}$
Complémentaire sur A^* : $\mathbf{Comp(L)} = A^* \backslash L = \{x \mid x \in A^* \text{ et } x \notin L\}$

Définition 1.17 - Opérateurs induits par la concaténation des mots

Produit : $\mathbf{LM} = \mathbf{L} \times \mathbf{M} = \{x \cdot y \mid x \in L \text{ et } y \in M\}$
Puissance : $\mathbf{L^0} = \{\varepsilon\}$ et $\mathbf{L^n} = L \times L^{n-1} = L^{n-1} \times L$
Fermeture de Kleene : $\mathbf{L}^* = \bigcup_{i=0\ldots\infty} L^i$
Fermeture positive : $\mathbf{L}^+ = \bigcup_{i=1\ldots\infty} L^i$

Propriétés des opérateurs entre langages

- Le langage vide est **absorbant** pour la concaténation des langages : $\emptyset \times L = \emptyset = \emptyset \times L$
- $< \mathcal{P}(A^*), \times, \{\varepsilon\} >$ est un monoïde :
  - Le langage neutre est élément neutre pour la concaténation des langages : $\{\varepsilon\} \times L = L = L \times \{\varepsilon\}$
  - La concaténation des langages est associative : $(L1 \times L2) \times L3 = L1 \times (L2 \times L3)$
- $L^+ = L \times L^* = L^* \times L$ et $L^* = \{\varepsilon\} \cup L^+$
- $\emptyset^* = \{\varepsilon\}$ et $\{\varepsilon\}^* = \{\varepsilon\}$

Exercice sur les langages et opérations

Montrer que le produit de deux langages préfixes est un langage préfixe

Exercice sur les langages et opérations - Démonstration

Montrer que le produit de deux langages préfixes est un langage préfixe

Démonstration : par l'absurde, assez simple en utilisant le lemme de Levi

- Soient L et M deux langages préfixes, $w_1, w_2 \in LM$
- **Démonstration par l'absurde** : supposons que w_1 est préfixe propre de $w_2 \Rightarrow \exists z$ tel que $w_1 \cdot z = w_2$
- De plus, $\exists u_1, u_2 \in L$ et $\exists v_1, v_2 \in M$ tels que $w_1 = u_1 \cdot v_1$ et $_2 = u_2 \cdot v_2$
- **Lemme de levi** sur $u_1 \cdot (v_1 \cdot z) = u_2 \cdot v_2$
  $\Rightarrow \exists! z$ tel que $(u_1 = u_2 \cdot w$ et $v_2 = w \cdot v_1 \cdot z)$ ou $(u_2 = u1 \cdot w$ et $v_1 = w \cdot v_2 \cdot z)$
- Or u_1 n'est pas préfixe de u_2 et réciproquement $\Rightarrow u_1 = u_2$ et $w = \varepsilon \Rightarrow v2 = v1 \cdot z$ ou $v1 = v2 \cdot z$
- **Contradiction** car v_1 n'est pas préfixe de v_2 et réciproquement
- **La supposition n'est pas vraie** : w_1 n'est pas préfixe propre de w_2
- De même w_2 ne peut pas être préfixe propre de $w_1 \Rightarrow LM$ est un langage préfixe