THÉORIES DES LANGAGES

Mr,HEMIOUD
hemourad@yahoo,fr
Université de Jijel
Département d'informatique

Notions fondamentales en Théorie des Langages

INTRODUCTION

- La théorie des Langages est une branche fondamentale de l'Informatique théorique.
- la première question que vous allez poser :
  - **Pourquoi faut-t-il étudier la théorie des Langages ?**
  - **C'est quoi l'apport pour un informaticien ?**

→Pour répondre à vos questions nous commençons ce chapitre par exposer des applications de cette théorie.

Exemple 1 : Correction d’orthographe et de grammaire dans un document Word

- Comment fonctionne un logiciel de correction d’orthographe ?
- Comment fonctionne un logiciel de correction de grammaire ?
- Si vous êtes entrain d’écrire un texte sur word, comment un mot peut être souligné en rouge ou en vert ?
- Quand est ce qu’un mot n’est pas souligné ?

Exemple2:Compilation d’un code avec le langage C

- Que se passe-t-il lorsque vous cliquez sur compiler ?
- Comment un compilateur vous présente les erreurs ?
- Le compilateur accepte tous les mots ?

INTRODUCTION

- Je suppose que vous posez déjà la question :

Y a-t-il une solution algorithmique pour ce genre de problèmes ?

→ la Théorie des Langages propose des solutions algorithmiques pour résoudre ces questions :

- **Y a-t-il une solution algorithmique permettant de vérifier : si un mot appartient à un Langage ?**

  → Pour répondre à cette question nous pouvons faire recours aux **automates**.

- **Y a-t-il une solution algorithmique permettant de vérifier : si une phrase respecte les règles de grammaire ?**

  →Pour répondre à cette question nous pouvons faire une étude des règles de **grammaire.**

LES LANGAGES

Définitions, propriétés et notations

Plusieurs notions sont fondamentales pour pouvoir présenter la théorie des langages.

- **Définition 1 (Alphabet)** : Un alphabet (*vocabulaire)*, noté A, est un ensemble fini non vide de symboles (lettres, de chiffres et/ou de symboles graphiques)
  - ***Exemples d'alphabets :***
    - l'alphabet du langage des nombre binaires : A= {0, 1}
    - Langage C (quel est son alphabet ?) ;
    - Langue française (quel est son alphabet ?).

- **Définition 2 (Mot)** : Un **mot** (ou bien une chaîne), défini sur un alphabet A, est une suite finie de symboles juxtaposés de A.
  - ***Exemples de mots :***
    - Sur l'alphabet A={a,b}, on peut construire les mots: a,aa, aba bba,….
    - Sur l'alphabet A={0,1}, on peut construire les mots: 1,10, 110 101,….
    - Sur l'alphabet A={a,b,(,),+}, on peut construire les mots: a,a+a, (a)=ba, b++,….
- **Définition 3 (Longueur d'un mot)** : La longueur d'un mot ***u*** défini sur un alphabet A, notée **|*u*|,** est le nombre de symboles qui composent ***u***.
  - La cardinalité d'un mot ***u*** par rapport à un symbole $a \in$ A (notée par $|u|_a$) est le ***nombre d'occurrence*** de a dans ***u***.
  - ***Par exemple :***

– $|abc| = 3$, $|1010111| = 7$.

– *Le mot dont la longueur est nulle est noté par* ϵ : $|\epsilon| = 0$.

– $|abc|_c = 1$, $|1010111|_1 = 5$, $|1010111|_0 = 2$.

- **Définition 4 (Mot vide)** : le mot vide, noté **ε**, est défini sur tous les alphabets et est le mot de longueur 0 (autrement dit, $\mathbf{|\varepsilon| = 0}$).

- **Définition 5 (A^+)** : on note A^+ l'ensemble des mots de *longueur* supérieure ou égale à 1 que l'on peut construire à partir de l'alphabet A.

- **Définition 6 (A^*)** : on note $\boldsymbol{A^*}$ l'ensemble des mots que l'on peut construire à partir de A, y compris le mot vide : $A^* = \{\varepsilon\} \cup A^+$
  - ***Exemple*** : Soit V = {a, b} alors :

    A* = {ε, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, bab, bba, …}

- **Concaténation des mots**

Soient $\boldsymbol{u}$ et $\boldsymbol{v}$ deux mots définis sur l'alphabet $\mathcal{A}$. La concaténation de $\boldsymbol{u}$ avec $\boldsymbol{v}$ est un mot $\boldsymbol{w}$ défini sur le même alphabet $\mathcal{A}$. $\boldsymbol{w}$ est obtenu en écrivant $\boldsymbol{u}$ suivi de $\boldsymbol{v}$:

  - $\boldsymbol{u} = \boldsymbol{a}_1 \ldots \boldsymbol{a}_n$, $\boldsymbol{v} = \boldsymbol{b}_1 \boldsymbol{b}_2 \ldots \boldsymbol{b}_m$
  - $\boldsymbol{w} = \boldsymbol{uv} = \boldsymbol{a}_1 \ldots \boldsymbol{a}_n \boldsymbol{b}_1 \boldsymbol{b}_2 \ldots \boldsymbol{b}_m$

- La concaténation est notée par le point, mais il peut être omis s'il n'y a pas de d'ambiguïté
- On écrira alors : $\boldsymbol{w} = \boldsymbol{u}.\boldsymbol{v} = \boldsymbol{uv}$
  - ***Exemples :***

Propriété de la concaténation

Soient $\boldsymbol{w}$, $\boldsymbol{u}$ et $\boldsymbol{v}$ trois mots définis sur l'alphabet A:

- $|\boldsymbol{u}.\boldsymbol{v}| = |\boldsymbol{u}| + |\boldsymbol{v}|$;
- $\forall \boldsymbol{a} \in A : |\boldsymbol{u}.\boldsymbol{v}|_{\boldsymbol{a}} = |\boldsymbol{u}|_{\boldsymbol{a}} + |\boldsymbol{v}|_{\boldsymbol{a}}$;
- $(\boldsymbol{u}.\boldsymbol{v}).\boldsymbol{w} = \boldsymbol{u}.(\boldsymbol{v}.\boldsymbol{w})$ (la concaténation est associative) ;
- $\boldsymbol{w}.\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}.\boldsymbol{w} = \boldsymbol{w}$ ($\boldsymbol{\varepsilon}$ est un élément neutre pour la concaténation) ;

- **L'exposant**

L'opération $\boldsymbol{w.w}$ est notée par $\boldsymbol{w^2}$. En généralisant, on note $\boldsymbol{w}^{\mathrm{n}} = \boldsymbol{w}...\boldsymbol{w}$ ($\boldsymbol{n}$ fois)

En particulier, l'exposant **0** fait tomber sur $\boldsymbol{\varepsilon}$: $\boldsymbol{w}^0 = \boldsymbol{\varepsilon}$ (le mot $\boldsymbol{w}$ est répété **0** fois).

- **Le mot miroir**

Soit $\boldsymbol{w = a_1a_2...a_n}$ un mot sur A. On appelle mot ***miroir*** de $\boldsymbol{w}$ et on le note par $\boldsymbol{w}^{\mathbf{R}}$ le mot obtenu en écrivant $\boldsymbol{w}$ l'envers, c'est-à-dire que $\boldsymbol{w^{\mathrm{R}} = a_n...a_2a_1}$. Il est donc facile de voir que $\boldsymbol{(w^{\mathrm{R}})^{\mathrm{R}} = w}$.

  - Certains mots, appelés ***palindromes***, sont égaux à leur miroir ($\boldsymbol{w}^{\mathbf{R}}= \boldsymbol{w}$). Par ailleurs, on peut facilement vérifier que : $\boldsymbol{(u.v)^{\mathrm{R}} = v^{\mathrm{R}}.u^{\mathrm{R}}}$.

Définition (Préfixe, suffixe et facteur) :

Soient deux mots $\boldsymbol{u}$ et $\boldsymbol{v}$ définis sur un alphabet A.

- $\boldsymbol{u}$ est un **préfixe** de $\boldsymbol{v}$ si et seulement si $\exists \boldsymbol{w} \in \mathbf{A^*}$ tel que: $\boldsymbol{uw} = \boldsymbol{v}$ **;**
- $\boldsymbol{u}$ est un **suffixe** de $\boldsymbol{v}$ si et seulement si $\exists \boldsymbol{w} \in \mathbf{A^*}$ tel que: $\boldsymbol{wu} = \boldsymbol{v}$ **;**
- $\boldsymbol{u}$ est un **facteur** de $\boldsymbol{v}$ si et seulement si $\exists \boldsymbol{w_1} \in \mathbf{A^*}$, $\exists \boldsymbol{w_2} \in \mathbf{A^*}$ tels que: $\boldsymbol{w_1 u w_2} = \boldsymbol{v}$**.**

Mots conjugués

Deux mots $\boldsymbol{x}$ et $\boldsymbol{y}$ sont dits conjugués s'il existe deux mots $\boldsymbol{u}$ et $\boldsymbol{v}$ tels que : $\boldsymbol{x} = \boldsymbol{uv}$ et $\boldsymbol{y} = \boldsymbol{vu}$**.**

Exemple

Soit le mot $x = ((acbc)^{\mathrm{R}}.baca)^{\mathrm{R}}$

- Donner la chaîne de caractères à laquelle x
- Quelle est la valeur de $|x|$, $|x|_{\boldsymbol{a}}$, $|x|_{\boldsymbol{b}}$ et $|x|_{\boldsymbol{c}}$?
- Donner un ***préfixe*** de $\boldsymbol{x}$ contenant au moins deux lettres '$\boldsymbol{c}$'.
- Donner un ***suffixe*** de $\boldsymbol{x}$ contenant une seule lettre '$\boldsymbol{a}$'.

Notions sur les langages

NOTIONS SUR LES LANGAGES

Un **langage** sur un alphabet A est un ***ensemble*** de mots construits sur A. Tout langage défini sur A est donc une partie de ***A****. On distingue en particulier le langage vide ϕ (*il ne contient aucun mot*)

Exemple :

- Langage des nombre binaires définies sur l'alphabet {0, 1} (infini) ;
- Langage des mots de longueur 2 défini sur l'alphabet {a, b}={aa, ab, ba, bb} ; (fini)
- Langage C (quel est son alphabet ?) ;
- Langue française (quel est son alphabet ?).

Description d'un langage :

- Un langage ***fini*** peut être décrit par *l'énumération* des mots qui le composent.
- Certains langages ***infinis*** peuvent être décrits par *l'application d'opérations* à des langages plus simples.
- Certains langages ***infinis*** peuvent être décrits **par un ensemble de règles appelé grammaire**
- Enfin, certains langages infinis ne peuvent pas être décrits, ni par l'application d'opérations, ni par un ensemble de règles. On parle alors de langage ***indécidable***.

Exemples : donner des exemples de mots contenus dans chacun des langages

1. Le langage des nombres binaires définies sur l'alphabet {0, 1}
2. Le langage des mots de longueur 2 définis sur l'alphabet {0, 1} = {00, 01, 10, 11}

Les langages suivants sont définies sur l'alphabet {a, b}

3. Le langage de tous les mots *commençant* par ***a***
4. L= {w∈{a,b}*, tel que ***w*** contient seulement 2b, le reste c'est des a's}
4. L = {w∈{a,b}*, tel que ***w*** contient un nombre de ***a*** divisible par 3}
5. L = {w∈{a,b}*, tel que ***w*** contient un nombre *impaire* de b}
6. L= {w∈A*, w contient 2***b*** consécutifs}
7. le langage des mots formés de ***n*** fois la lettre ***a*** suivi de ***n*** fois la lettre ***b***
8. Le langage des mots contenant au moins une fois la lettre ***a***

Opérations sur les langages:

Soient L, L_1 et L_2 trois langages définis sur l'alphabet A, nous définissons les opérations suivantes :

- **L'union** : notée par + ou | plutôt que $\cup$.

$$L_1 \,|\, L_2 = L_1 + L_2 = \{m \textit{ tel que } m \in L_1 \vee m \in L_2\} \;;$$

- **L'intersection** :

$$L_1 \cap L_2 = \{m \textit{ tel que } m \in L_1 \wedge m \in L_2\} \;;$$

- **Le complément** :

$$L = \{\text{tous les mots } m \text{ sur A } \textit{tel que } m \notin L\} \;;$$

- La **concaténation** (opération non ensembliste) :

$$L_1.L_2 = \{m \textit{ tel que } \exists u \in L_1, \exists v \in L_2 : m = uv\} \;;$$

- **Exposant** (opération non ensembliste) :

$$L^{\boldsymbol{n}} = L.L \,.\,.\,.\, L \;(\, n \textit{ fois} \,)$$

$$= \{m \textit{ tel que } \exists u_1, u_2, \ldots u_n \in L : m = u_1 u_2 \ldots u_n\}$$

- **Fermeture transitive de Kleene** (opération non ensembliste) : notée $L^* = \cup_{i \geq 0} L^i$.

En particulier, si L = A on obtient A*: l'ensemble de tous les mots possibles sur l'alphabet A.

On peut ainsi définir un langage comme étant un sous-ensemble quelconque de A;

- **Fermeture non transitive** (opération non ensembliste) :

$$L^+ = \cup_{i>0} L^i \ ;$$

- **Le langage miroir** (opération non ensembliste) :

$$L^R = \{m \text{ tel que } \exists u \in L : m = u^R\}$$

Propriétés des langages

Soit L, L_1, L_2, L_3 quatre langages définis sur l'alphabet A :

- $L^0 = \{\epsilon\}$
- $L^1 = L$
- $L^2 = L.L$
- $L^{n+1} = L.L^n = L^n.L$
- $L^* = L^+|\{\epsilon\}$;
- $L_1.(L_2.L_3) = (L_1.L_2).L_3$;
- $L_1.(L_2|L_3) = (L_1.L_2)|(L_1.L_3)$;
- $L.L \neq L$;
- $L_1.(L_2 \cap L_3) \neq (L_1 \cap L_2).(L_1 \cap L_3)$;
- $L_1.L_2 \neq L_2.L_1$;
- $(L^*)^* = L^*$;
- $L^*.L^* = L^*$;
- $L_1.(L_2.L_1)^* = (L_1.L_2)^*.L1$;
- $(L_1|L_2)^* = (L_1^* L_2^*)^*$;
- $L_1^*|L_2^* \neq (L_1|L_2)^*$;

- L'application des divers opérateurs doit respecter des priorités bien définies comme suit : | , . , *