le cnam

Codes et Automates finis

Cours 8 Langages rationnels

- Introduction

En 1936, Alan Turing [1912–1954, mathématicien britannique] présente sa “machine de Turing”, c’est à dire imagine un procédé mécanique pour mettre en œuvre un algorithme. Il s’agit d’une machine à mémoire, ou “à états”, qui interagit avec un texte. D’un point de vue pratique, la machine est composée des éléments suivants : une bande infinie découpée en cases, une tête de lecture et écriture qui permet de modifier le contenu du ruban, un alphabet qui contient le symbole “blanc”, un ensemble d’états, noté Q, avec en particulier un état initial $q_0 \in Q$ et un ou plusieurs états finals $F \subset Q$, un programme, c’est à dire un ensemble d’instructions pour la machine.

Très vite, Alonzo Church [1903–1995, mathématicien américain] se rend compte que la machine de Turing est universelle. La “thèse de Church” exprime que tout problème de calcul fondé sur une procédure algorithmique peut être résolu par une machine de Turing.

Le fait de commander une machine avec un programme a fait émerger une nouvelle branche de la linguistique. Loin de se désintéresser des langues naturelles qui permettent la communication entre les êtres humains, il convient aussi d’étudier et définir les langues artificielles qui permettent de donner des ordres à une machine. Une langue artificielle se compose d’un alphabet (un ensemble fini de lettres), un vocabulaire ou dictionnaire des mots courants, une grammaire qui énonce les règles de production de nouveaux mots, comme les règles du pluriel ou la conjugaison dans le cas des langues naturelles. Au delà de ces règles de syntaxe, la sémantique, qui donne un sens, une signification aux mots et aux phrases, reste une des grandes caractéristiques des langues naturelles.

En 1959, Noam Chomsky [né en 1928, linguiste américain] a proposé une hiérarchie entre les grammaires selon le type de règles que l’on peut utiliser. Il a aussi proposé quatre classes principales qui vont du plus complexe au plus élémentaire : langage récursivement énumérable, langage contextuel, langage algébrique, langage rationnel. Les langages rationnels sont les plus simples. L’objet de cette seconde partie du cours est de donner une introduction envue de leur étude et leur utilisation.

Entretemps Stephen Kleene [1909–1994, mathématicien américain] a établi le “théorème de Kleene” entre un langage rationnel et un “automate fini”. On peut voir un automate fini comme une machine de Turing très simplifiée. Il est défini par un quintuple : un ensemble fini d’états Q, un alphabet A, fini lui aussi, une fonction de transition δ qui, à tout couple (q,a) formé d’un état $q \in Q$ et d’une lettre $a \in A$, associe un nouvel état $q' = \delta(q,a)$; on a donc $Q \times Q(q,a) \longmapsto q' = \delta(q,a) \in Q$, un état initial $q_0 \in Q$ (et parfois plusieurs !), un ensemble F d’états finals : $F \in Q$.

La relation entre langage rationnel et automate fini constitue l'ossature de cette seconde partie du cours. Plus généralement, on peut placer en face de chaque type de langage de la hiérarchie de Chomsky une famille d'automates (ou de machines) :

$$\begin{array}{rcl} \text{langage récursivement énumérable} & \longleftrightarrow & \text{machine de Turing} \\ \text{langage contextuel} & \longleftrightarrow & \text{automate linéairement borné} \\ \text{langage algébrique} & \longleftrightarrow & \text{automate à pile} \\ \text{langage rationnel} & \longleftrightarrow & \text{automate fini} \end{array}$$

Le "lemme d'Arden" (1961) [Dean Arden, 1925–2018, informaticien américain] permet la résolution d'équations pour les langages rationnels.

Parmi les exemples d'automates finis dans la vie courante de l'utilisateur des systèmes automatiques et de l'informatique, citons pêle–mêle un digicode, un bouton de commande d'ascenseur, un distributeur automatique de boissons, un protocole de communication ou un éditeur de texte dans la fonction de recherche d'une chaîne de caractères, comme par exemple la commande "grep" du système d'exploitation Unix. Rappelons pour terminer que la mémoire d'un "automate fini" est limitée au nombre fini des états, c'est à dire au nombre d'éléments de l'ensemble Q.

- Définitions

Un alphabet A est un ensemble non vide fini ; ses éléments sont appelés "lettres", ou "symboles". Par exemple, $A = \{a, b\}$.

Un mot sur un alphabet A est une suite finie de lettres de A : Si n est un nombre entier ≥ 1 et $w_1, w_2, \ldots, w_n$ est une suite de lettres, $w = w_1 w_2 \ldots w_n$ est un mot sur l'alphabet A. Par exemple, a, b, ab, ba, ... sont des mots sur l'alphabet $A = \{a, b\}$.

La longueur $|w|$ du mot $w = w_1 w_2 \ldots w_n$ est le nombre entier n supérieur ou égal à 1 qui désigne le nombre de lettres du mot w.

Le "mot sans lettre" ou "mot vide" ε est le mot de longueur nulle. Noter que que le mot sans lettre n'appartient pas à l'alphabet : $\varepsilon \notin A$.

Un langage L sur un aphabet A est un ensemble de mots qui utilisent des lettres de l'alphabet A. Par exemple, l'ensemble des mots de deux lettres sur l'alphabet $A = \{a, b\}$ constitue le langage $L = \{aa, ab, ba, bb\}$. Un langage peut être égal à l'ensemble vide ($L = \emptyset$) et il ne comporte alors aucun mot. Il peut être égal au mot sans lettre ($L = \{\varepsilon\}$). Il peut comporter aussi un nombre infini de mots.

L'ensemble de tous les mots, y compris le mot vide, formés sur l'alphabet A est noté A^*. Par exemple, pour $A = \{a, b\}$, on a $A^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$.

Un abus très courant consiste à confondre la lettre de l'alphabet a ($a \in A$) et le mot d'une lettre $w = a$.

- Concaténation

Le produit de concaténation, ou simplement concaténation, $u.v$ noté aussi uv des mots $u = u_1 u_2 \ldots u_n$ et $v = v_1 v_2 \ldots v_m$ est le mot $w = uv$ défini par $w = u_1 u_2 \ldots u_n v_1 v_2 \ldots v_m$. Par exemple, $a.ab = aab$, $ab.a = aba$. On voit sur cet exemple très simple que la concaténation n'est pas une opération commutative : on a en général $uv \neq vu$.

On étend la définition précédente au mot sans lettre : $\varepsilon u = u\varepsilon = u$ pour un mot u quelconque. Le mot sans lettre est donc un élément neutre pour la concaténation. C'est l'analogue du nombre "1" pour la multiplication.

- Préfixe

On dit que le mot u est un préfixe du mot w si et seulement si il existe un mot v de sorte que $w = uv$. Par exemple, a est un préfixe du mot aba et ab est également un préfixe du mot aba.

- Monoïde

On peut présenter la notion de concaténation comme une loi de composition sur l'ensemble A^* de tous les mots.
Par définition, un monoïde est la donnée $(M, .)$ d'un ensemble M et d'une loi de composition : pour tout $u, v \in M$, le produit $u.v$ est un nouvel élément de M. On suppose de plus que le produit . est associatif : $u.(v.w) = (u.v).w$, $\forall u, v, w \in M$ et qu'il existe un élément neutre $\varepsilon \in M$: $u.\varepsilon = \varepsilon.u = u$, $\forall u \in M$.
On peut démontrer facilement que le produit de concaténation est associatif. L'ensemble A^* de tous les mot sur l'alphabet fini A, muni du produit de concaténation, est un monoïde ; c'est le "monoïde libre" engendré par l'alphabet A.

- Union, ou somme

Nous abordons maintenant les opérations dites "opérations rationnelles", sur les langages.
Attention, un langage $L \subset A^*$ est un ensemble de mots. Chaque mot est la concaténation de lettres de l'alphabet A.
On se donne deux langages L et L' qui utilisent les lettres de l'alphabet A. Leur somme $L+L'$ est simplement leur réunion ensembliste : $L+L' = L \cup L'$.
Par exemple, pour $A = \{a, b\}$, $L = \{a, ab, ba\}$ et $L' = \{b, ba, aba\}$, on a
$L+L' = \{a, b, ab, ba, aba\}$.
Autre exemple : $\{\varepsilon, a, b\} + \{b, ab, ba\} = \{\varepsilon, a, b, ab, ba\}$.

- Concaténation, ou produit

La concaténation d'un langage L composé avec des lettres de l'alphabet A ($L \subset A^*$) et du langage L' également composé avec des lettres de l'alphabet A ($L' \subset A^*$) est par définition le langage $L.L'$ (noté parfois plus simplement LL') obtenu en faisant la concaténation de tous les mots de L avec tous les mots de L': $L.L' = \{uv, u \in L, v \in L'\}$.
Par exemple, $\{\varepsilon, a, ab\}.\{b, ab, ba\} = \{b, ab, ba, aab, aba, abb, abab, abba\}$.
Il est aussi bien utile d'adopter la notation exponentielle $u^n = u.u\ldots.u$ avec le produit répété n fois. Ainsi $u^2 = u.u$, $u^3 = u.u.u$, *etc*. On adopte aussi la convention $u^0 = \varepsilon$.
On peut réécrire le produit précédent sous la forme
$\{\varepsilon, a, ab\}.\{b, ab, ba\} = \{b, ab, ba, a^2b, (ab)^2, ab^2a\}$.

- Associativité du produit de concaténation des langages

On se donne un alphabet A et trois langages K, L et M sur cet alphabet : $K \subset A^*$, $L \subset A^*$, $M \subset A^*$. On rappelle que $K.L = \{uv, u \in K, v \in L\}$. Alors $K.(L.M) = (K.L).M$: le produit de concaténation des langages est associatif.

La preuve de cette proposition est une conséquence directe du fait que la concaténation des mots est une opération associative. On a en effet $u(vw) = (uv)w$ pour tout $u \in K,\ v \in L,\ w \in M$. Donc $K.(L.M) = \{u(vw), u \in K, v \in L, w \in M\} = \{(uv)w, u \in K, v \in L, w \in M\} = (K.L).M$.

- Distributivité de la concaténation par rapport à la somme

Pour trois langages quelconques K, L, M sur un alphabet A, on a $K.(L+M) = (K.L)+(K.M)$. En effet, $K.(L+M) = K.(L \cup M) = \{u.w, u \in K, w \in L \cup M\} = \{u.w, u \in K, w \in L \text{ ou } w \in M\}$ $= \{u.w, u \in K, w \in L\} \cup \{u.w, u \in K, w \in M\} = (K.L) \cup (K.M) = (K.L)+(K.M)$.
On a la même propriété pour le produit de concaténation à droite de la somme :
$(K+L).M = (K.M)+(L.M)$.

- Étoile, appelée parfois "étoile de Kleene"

On se donne un langage L sur l'alphabet A : $L \subset A$. On pose $L^0 = \{\varepsilon\}$, $L^1 = L$ $L^2 = L.L$ et de façon générale pour tout entier $i \geq 0$, $L^{i+1} = L^i.L$. L'étoile L^* réunit cette suite infinie de langages : $L^* = \cup_{i \geq 0} L^i = \{\varepsilon\} \cup L \cup L^2 \cup L^3 \cup \ldots$.
Si on adopte un langage mathématique, L^* est le plus petit (au sens de l'inclusion) monoïde contenant L ; on parle aussi de monoïde libre engendré par L.
Si $\emptyset$ désigne l'ensemble vide, on a $\emptyset^* = \{\varepsilon\}$. La notation A^* désigne à la fois l'opérateur étoile appliqué à l'alphabet A et l'ensemble de tous les mots (y compris le mot vide ε) qui utilisent l'alphabet A. En effet, pour i entier positif ou nul, A^i est le langage formé des mots de longueur i avec les lettres de A et on a bien $A^* = \{\varepsilon\} \cup A \cup A^2 \cup A^3 \cup \ldots$.
Considérons par exemple $L = \{a, ba\}$ sur l'alphabet $A = \{a, b\}$. On a
$L^2 = \{a^2, aba, baa, baba\}$, $L^3 = \{aaa, aaba, abaa, baaa, babaa, baaba, ababa, bababa\}$, ...
On se rend compte que L^* est l'ensemble de tous les mots où chaque "b" est suivi d'un "a".
Autres exemples :

A^*a est l'ensemble des mots qui se terminent par "a",
a^* l'ensemble des mots qui ne contiennent que la lettre "a" ainsi que le mot sans lettre ε,
$a^* + b^*$ l'ensemble des mots ne contenant que des"a" ou que des "b" (et ε),
$(AA)^*$ l'ensemble des mots de longueur paire (ou nulle !),
A^*aA^* l'ensemble des mots contenant au moins une fois une occurence de la lettre "a".

- "Étoile tronquée"

On se donne un langage $L \subset A$. On pose $L^+ = LL^* = L^*L = L \cup L^2 \cup L^3 \cup \ldots = \cup_{i \geq 1} L^i$.
Si le langage L contient le mot vide, alors on a $L^+ = L$.
Si le langage L ne contient pas le mot vide, on a $L^+ = L^* \setminus \{\varepsilon\}$. Dans ce cas, le langage L^+ est obtenu à partir du langage L^* en enlevant le mot sans lettre ε. Par exemple, pour un alphabet A (qui par définition ne contient pas le mot vide !), l'ensemble A^+ est constitué des mots non vides.

Exercices

- Langages

Sur l'alphabet $A = \{a, b\}$, on définit les langages suivants : $L_1 = \{b, ab\}$, $L_2 = \{ba\}$, $L_3 = \{\varepsilon, ab\}$.

a) Déterminer les langages suivants : $L_1 + L_2 + L_3$, $L_1 L_2$, $L_1 L_3$, $(L_3)^2$, $L_1 (L_2 + L_3)$.

b) Comparer les langages $(aL_2)^*$ et $(L_3 a)^*$.

c) Quel est le plus petit entier n tel que le mot $u = aabbab$ appartienne à $(a + L_1)^n$?

- Étoile

Avec $A = \{a, b\} = a + b$, vérifier si on a égalité ou pas entre les couples de langages suivants

a) $(a+b)^*$ et $a^* + b^*$

b) $(aba)^*$ et $(ab^*a)^*$

c) $ab^+ + a$ et ab^*, avec $b^+ = b + b^2 + b^3 + \cdots$

- Identités

On se donne un langage $L \subset A$; on rappelle que $L^+ = L \cup L^2 \cup L^3 \cup \ldots = \cup_{i \geq 1} L^i$. Montrer que

a) $(L^*)^* = L^*$

b) Si $\varepsilon \in L$, alors $L^+ = L^*$

c) Si $\varepsilon \notin L$, alors $L^+ = L^* \setminus \{\varepsilon\}$.

d) Si $\varepsilon \notin L$, alors $L^+ = LL^* = L^* L$.

e) $(L^+)^+ = L^+$

f) Si $\varepsilon \notin L$, alors $L^* L L^* = L^+$.

g) $L^* = LL^* + \{\varepsilon\} = L^* L + \{\varepsilon\}$