

TD n°2

La reconnaissance de motifs

L'objet du problème est de montrer que le modèle des automates finis conduit naturellement à la construction d'algorithmes efficaces pour la reconnaissance de motifs dans un texte. La reconnaissance de motifs est une fonction essentielle des logiciels de traitement de texte.

Rappels et notations

On appelle *alphabet* un ensemble fini A . Les éléments de A sont appelés des *lettres*. On note A^* l'ensemble des suites finies d'éléments de A qui sont appelées des *mots* sur A ou simplement des mots lorsqu'aucune ambiguïté n'est à craindre. Le mot vide est noté 1_{A^*} .

Soient $f, g \in A^*$, on note fg le mot obtenu en faisant suivre la suite des lettres de f par la suite des lettres de g , $fg \in A^*$. Soit $f \in A^*$, s'il existe $g, h \in A^*$ tels que $f = gh$ alors g est un préfixe de f et h est un suffixe de f . Si $h \neq f$, h est dit un *suffixe propre* de f et g est dit un *préfixe non-vide* de f . S'il existe $u, v, w \in A^*$ tels que $f = uvw$, v est un facteur de f . Enfin, on note $|f|$ la longueur de f , c'est-à-dire le nombre de lettres de l'alphabet qui le composent.

Un *automate* est une structure $\langle Q, A, E, I, T \rangle$ telle que Q est un ensemble fini d'éléments appelés *états*, A est un alphabet, $I \subseteq Q$, les éléments de I sont appelés *états initiaux*, $T \subseteq Q$, les éléments de T sont appelés *états terminaux*, $E \subseteq Q \times A \times Q$, les éléments de E sont appelés des *transitions*.

Un *automate* est représenté graphiquement par un graphe orienté de la manière suivante: chaque état e est représenté par un nœud du graphe, chaque transition (e, a, f) est représentée par une flèche étiquetée par le mot a et qui va de l'état e vers l'état f . Pour marquer qu'un état est initial on ajoutera une flèche entrante sans origine. Pour marquer qu'un état est terminal on ajoutera une flèche sortante sans but. La figure 1 illustre ce mode de représentation graphique.

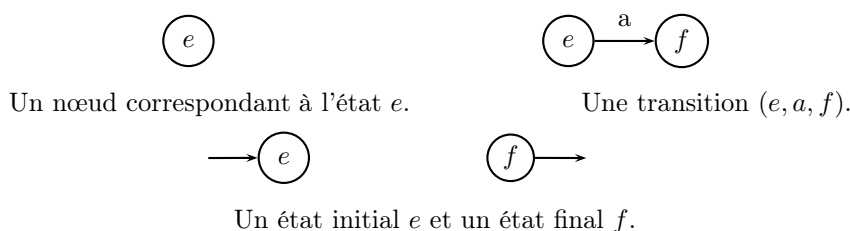
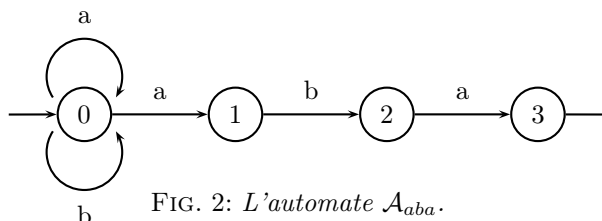


FIG. 1: Représentation graphique d'un automate.

Si pour chaque état $e \in Q$ et pour chaque lettre $a \in A$, il existe au plus un état $f \in Q$ tel que (e, a, f) soit une transition de l'automate, on dit que l'automate est déterministe. Dans ce cas, si $(e, a, f) \in E$ on choisit alors de noter f par $e.a$. On n'attachera pas d'importance au fait que la notation $e.a$ n'est pas définie pour tous les couples $(e, a) \in Q \times A$.

N.B. Dans tout le problème, l'alphabet A est fixé: $A = \{a, b\}$.

1. On considère l'automate $\mathcal{A}_{aba}$ décrit graphiquement par la figure 2.

FIG. 2: L'automate $\mathcal{A}_{aba}$.

(a) Décrivez le langage reconnu par l'automate $\mathcal{A}_{aba}$.

(b) Déterminez l'automate $\mathcal{A}_{aba}$. Donnez la table de transition de l'automate déterminisé ainsi que sa représentation graphique.

2. On considère l'automate $\mathcal{A}_{abaab}$ décrit graphiquement par la figure 3.

(a) Décrivez le langage reconnu par l'automate $\mathcal{A}_{abaab}$.

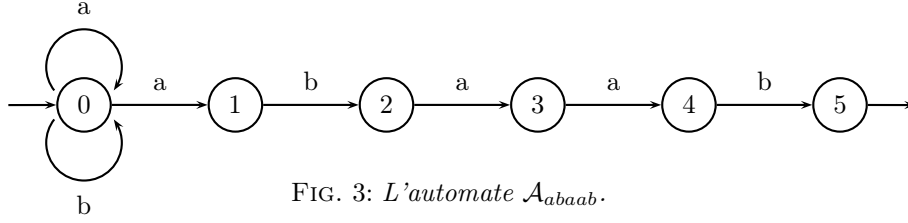


FIG. 3: L'automate $\mathcal{A}_{abaab}$.

- (b) Déterminez l'automate $\mathcal{A}_{abaab}$. Donnez uniquement la table de transition de l'automate déterminisé.

Dans la suite du problème, on considère un mot $f = a_1 \dots a_n \in A^*$ et l'automate $\mathcal{A}_f$ décrit graphiquement par la figure 4 page suivante.

On note P l'ensemble des préfixes du mot f :

$$P = \{1_{A^*}, a_1, a_1a_2, \dots, a_1 \dots a_{n-1}, a_1 \dots a_n\}$$

L'ensemble des états de l'automate $\mathcal{A}_f$ peut être identifié avec l'ensemble des préfixes de f : l'état 0 est identifié avec 1_{A^*} et pour $i \in [1, n]$, l'état i est identifié avec le préfixe $a_1 \dots a_i$. On note $\mathcal{D}_f$ le déterminisé de l'automate $\mathcal{A}_f$ obtenu par la méthode des sous-ensembles. L'ensemble des états de $\mathcal{D}_f$ est un sous-ensemble de l'ensemble des parties de P .

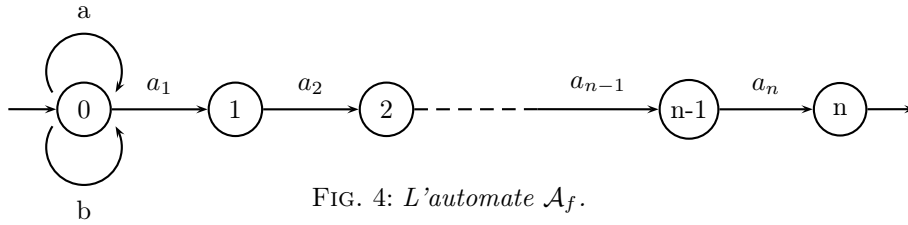


FIG. 4: L'automate $\mathcal{A}_f$.

3. Soit $\omega \in A^*$ un mot quelconque.

- Montrer que l'état $\mathcal{D}_f$ atteint à partir de l'état initial par la lecture de w est le sous ensemble de P constitué des suffixes de w .
- Montrez que cet ensemble est entièrement déterminé par son élément le plus long. Dans la suite du problème, on identifiera cet ensemble avec son élément le plus long.
- Déduisez de ce qui précède le nombre d'états de $\mathcal{D}_f$.

4. Soit α la fonction de $P \setminus \{1_{A^*}\}$ dans P qui associe à chaque préfixe non-vide p le plus long suffixe propre de p que est dans P . La fonction α n'est pas définie pour le mot vide 1_{A^*} .

- (a) Complétez la table 1 en donnant les valeurs de α dans le cas où f et le mot $abaab$. [h]

p	a	ab	aba	$abaa$	$abaab$
$\alpha(p)$					

TAB. 1: La fonction α pour $f = abaab$.

- (b) Montrez que les transitions de $\mathcal{D}_f$ sont bien définies récursivement par :

$$\forall x \in A, \begin{cases} 1_{A^*}.x &= x \text{ si } x \in P \\ &= 1_{A^*} \text{ sinon} \end{cases}$$

et

$$\forall x \in A, \forall p \in P \setminus \{1_{A^*}\}, \begin{cases} p.x &= px \text{ si } px \in P \\ &= \alpha(p) \text{ sinon} \end{cases}$$

5. Montrez que la fonction α elle-même peut être calculée récursivement par :

$$\forall x \in A, \alpha(x) = 1_{A^*}$$

et

$$\forall x \in A, \forall p \in P \setminus \{1_{A^*}\}, \begin{cases} \alpha(p.x) &= \alpha(p)x \text{ si } \alpha(p)x \in P \\ &= \alpha(\alpha(p)x) \text{ sinon} \end{cases}$$

6. Déduisez de ce qui précède un algorithme qui reconnaît toutes les occurrences d'un facteur f dans un mot t donné avec une complexité en $O(|f| + |t|)$.