Introduction à l'algorithmique de texte

Chaîne de caractères de longueur n sur un alphabet Σ.

Trouver dans cette chaîne les occurrences { d'un motif / d'un mot de l'ensemble des mots d'un dictionnaire

Applications

→ Extraction d'informations
→ Recherche dans un éditeur
→ Génomique, bioinformatique
→ Correction orthographique
→ ...

Déf. Un préfixe (strict) d'un mot $u \in \Sigma^*$ est un mot $v \in \Sigma^*$ t.q. il existe $w \in \Sigma^*$ (différent de u) avec $u = vw$.

Un suffixe d'un mot $u \in \Sigma^*$ est un mot $v \in \Sigma^*$ t.q. il existe $w \in \Sigma^*$ avec $u = wv$.

Chaîne de taille n dans laquelle on veut rechercher des motifs.

Motif	Recherche	Technique	Complexité de compilation	Complexité de recherche
e.r. de taille k	1 occurrence ou Hes les occurrences	AFD pour .*e (une occurrence à chaque passage par un état final	$O(\alpha^k)$	$O(n)$
chaîne de taille k	1 occ. / Hes les occ.	naïf	0	$O(n \times k)$
chaîne de taille k	Hes les occ.	Knuth-Morris-Pratt (KMP)	$O(k)$	$O(n)$
ensemble de m chaînes de taille k	correspondance exacte	Trie	$O(m \times k)$	$O(n)$
~~	~~~	Arbre de Patricia	$O(m \times k)$	$O(n)$ (plus compact que les Trie)
ensemble de m chaînes de taille k	Hes les occ.	m x KMP	$O(m \times k)$	$O(m \times n)$
ensemble de m chaînes de taille k	Hes les occ.	Aho-Corrasick (Trie + idée de KMP / transitions de repli)	$O(m \times k)$	$O(n$ + taille_du_résultat)

+ Arbre des suffixes (indexer la longue chaîne de caractères)

Motif : a b c a b d

Chaîne : a b a c a b c a b c a b d

a b

a

a b c [a b / a b c] a b d ← match

naïf

à optimiser

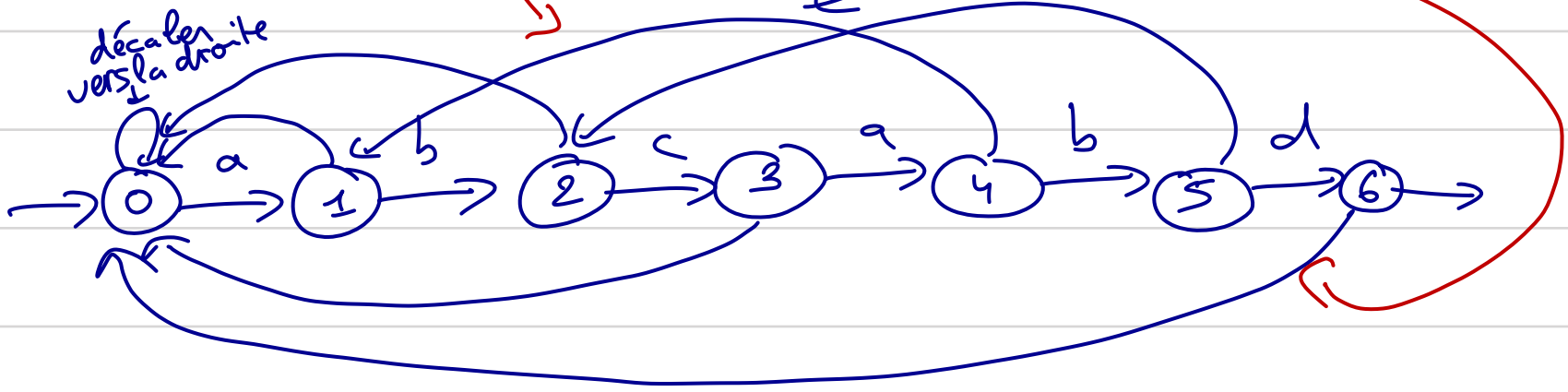
On se souvient, dans le motif, de la taille du plus long préfixe strict qui est un suffixe de chaque préfixe du motif.

Compilation

a b c a b d

(-1) 0 0 1 2 0

transition de replis



a b a c a b c a b c a b d

1 2
0 1
0 0 1 2 3 4 5
2 3 4 5 6

KMP

O(n)

replis ≤ n

O(k) par programmation dynamique

[a b c a b d

(-1) 0 ~~*~~

0 1 2 ~~3~~

~~*~~

0

Trie

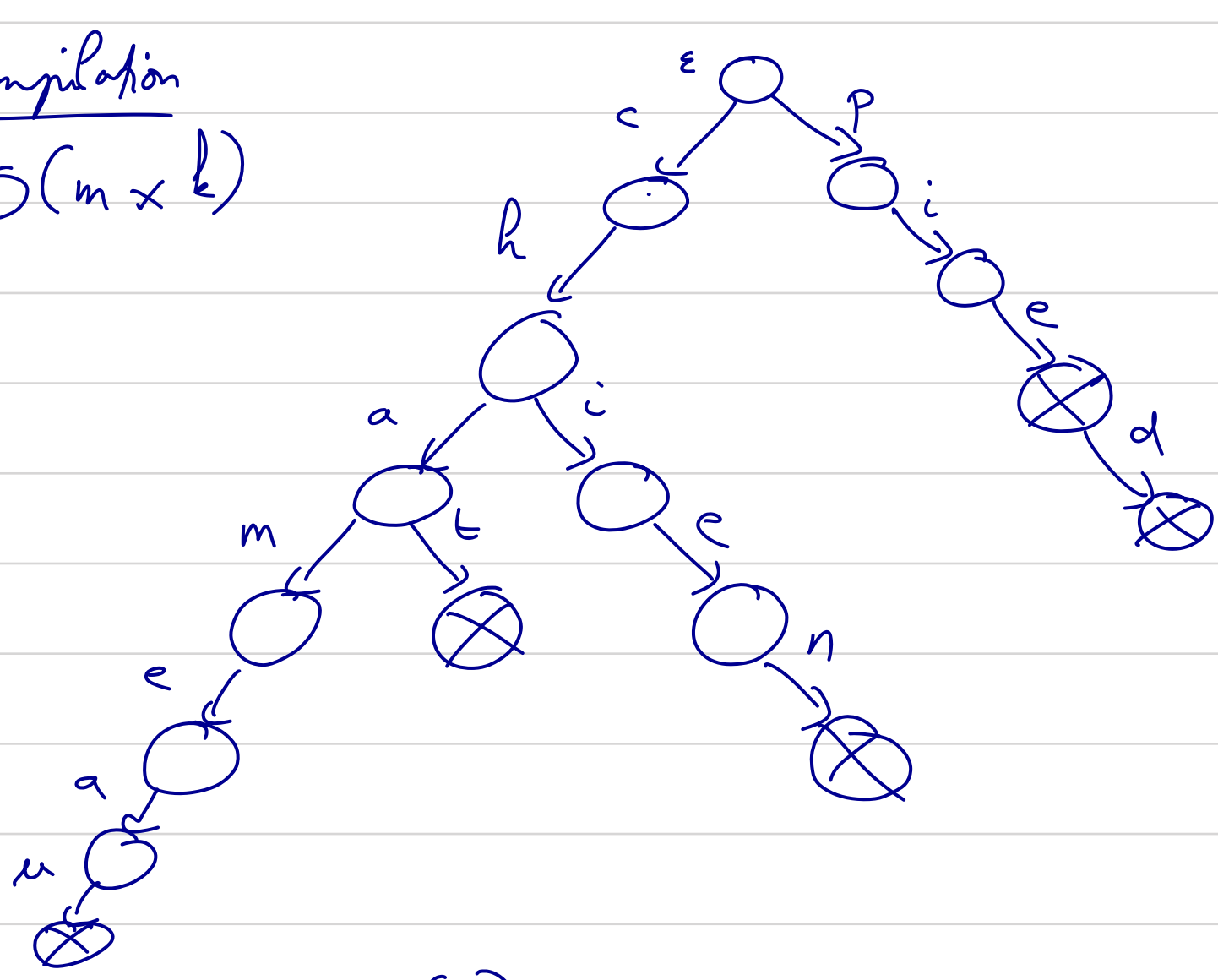
("retrieval")

Arbre dat chaque noeud correspond aux préfixes des mots stockés, chaque lien enfant-parent à un symbole, et où les noeuds sont marqués pour indiquer si ce préfixe est l'un des mots stockés.

{ chien, chat, chameau, pie, pied } $m = 5$ $k = 6$

Compilation

$O(m \times k)$



Recherche : $O(n)$

Arbres de Patricia : raffinement des Trie dans lequel les transitions parent-enfant peuvent être étiquetées par des sous-chaînes