2.3 Algorithme de Boyer-Moore

NSI Tle - JB Duthoit

2.3.1 Généralités

- Dû à Robert S. Boyer et J. Strother Moore – 1977
- utilisé le plus souvent dans les éditeurs de texte (tel quel ou optimisé)

2.3.2 Principe

L'algorithme de Boyer-Moore se base sur les caractéristiques suivantes :

- L'algorithme effectue un pré-traitement du motif. Cela signifie que l'algorithme "connait" les caractères qui se trouvent dans le motif
- On commence la comparaison motif-chaine par la droite du motif. Par exemple pour le motif CGGCAG, on compare d'abord le G, puis le A, puis C...on parcourt le motif de la droite vers la gauche
- Dans la méthode naïve, les décalages du motif vers la droite se faisaient toujours d'un "cran" à la fois. L'intérêt de l'algorithme de Boyer-Moore, c'est qu'il permet, dans certaines situations, d'effectuer un décalage de plusieurs crans en une seule fois.

Etape 1 :

CAAGCGCACAAGACGCGGCAGACCTTCGTTATAGGCGATGATTT
ACG
A et G ne correspondent pas. On sait qu'il y a un A dans le motif... On décale donc de deux crans vers la droite.

Etape 2 :

CAAGCGCACAAGACGCGGCAGACCTTCGTTATAGGCGATGATTT
  ACG
C et G ne correspondent pas. On sait qu'il y a un C dans le motif... On décale donc d'un cran vers la droite.

Etape 3 :

CAAGCGCACAAGACGCGGCAGACCTTCGTTATAGGCGATGATTT
    ACG

G et G se correspondent, ainsi que C et C. En revanche, G et A ne se correspondent pas...G n'est pas dans le "reste" de la clé...on décale donc de 3 crans vers la droite.

Etape 4 :

CAAGCGCACAAGACGCGGCAGACCTTCGTTATAGGCGATGATTT
      ACG

G et C ne correspondent pas. C est présent dans la clé, donc on décale d'un cran..

... et ainsi de suite jusqu'au moment ou l'on trouve (ou pas !) une correspondance parfaite.

Exercice 14.12
On considère le texte suivant :
ATAACAGGAGTAAATAACGGCTGGAGTA
et le motif CGGTG

1. Essayer de tête d'imaginer l'algorithme naïf et déterminer en combien d'étapes il trouvera l'indice de l'occurrence.
2. Détailler l'algorithme de Boyer Moore sur cet exemple. Combien d'étapes sont nécessaires ?

Exercice 14.13
On considère le texte suivant :
ABRACADABRA
et le motif DAB

1. Essayer de tête d'imaginer l'algorithme naïf et déterminer en combien d'étapes il trouvera l'indice de l'occurrence.
2. Détailler l'algorithme de Boyer Moore sur cet exemple. Combien d'étapes sont nécessaires ?

Remarque
Plus le motif est long, plus l'algorithme de Boyer Moore est efficace.

2.3.3 Pré-traitement

Principe

Le pré-traitement du motif consiste à garder dans une liste en mémoire la position de la dernière occurrence de chaque caractère distinct du motif (sauf le dernier , car il est utilisé pour comparer)

On utilise pour cela l'alphabet ASCII qui contient 256 caractères.

Dec	Bin	Hex	Char	Dec	Bin	Hex	Char	Dec	Bin	Hex	Char	Dec	Bin	Hex	Char
0	0000 0000	00	[NUL]	32	0010 0000	20	space	64	0100 0000	40	@	96	0110 0000	60	`
1	0000 0001	01	[SOH]	33	0010 0001	21	!	65	0100 0001	41	A	97	0110 0001	61	a
2	0000 0010	02	[STX]	34	0010 0010	22	"	66	0100 0010	42	B	98	0110 0010	62	b
3	0000 0011	03	[ETX]	35	0010 0011	23	#	67	0100 0011	43	C	99	0110 0011	63	c
4	0000 0100	04	[EOT]	36	0010 0100	24	$	68	0100 0100	44	D	100	0110 0100	64	d
5	0000 0101	05	[ENQ]	37	0010 0101	25	%	69	0100 0101	45	E	101	0110 0101	65	e
6	0000 0110	06	[ACK]	38	0010 0110	26	&	70	0100 0110	46	F	102	0110 0110	66	f
7	0000 0111	07	[BEL]	39	0010 0111	27	'	71	0100 0111	47	G	103	0110 0111	67	g
8	0000 1000	08	[BS]	40	0010 1000	28	(	72	0100 1000	48	H	104	0110 1000	68	h
9	0000 1001	09	[TAB]	41	0010 1001	29	)	73	0100 1001	49	I	105	0110 1001	69	i
10	0000 1010	0A	[LF]	42	0010 1010	2A	*	74	0100 1010	4A	J	106	0110 1010	6A	j
11	0000 1011	0B	[VT]	43	0010 1011	2B	+	75	0100 1011	4B	K	107	0110 1011	6B	k
12	0000 1100	0C	[FF]	44	0010 1100	2C	,	76	0100 1100	4C	L	108	0110 1100	6C	l
13	0000 1101	0D	[CR]	45	0010 1101	2D	-	77	0100 1101	4D	M	109	0110 1101	6D	m
14	0000 1110	0E	[SO]	46	0010 1110	2E	.	78	0100 1110	4E	N	110	0110 1110	6E	n
15	0000 1111	0F	[SI]	47	0010 1111	2F	/	79	0100 1111	4F	O	111	0110 1111	6F	o
16	0001 0000	10	[DLE]	48	0011 0000	30	0	80	0101 0000	50	P	112	0111 0000	70	p
17	0001 0001	11	[DC1]	49	0011 0001	31	1	81	0101 0001	51	Q	113	0111 0001	71	q
18	0001 0010	12	[DC2]	50	0011 0010	32	2	82	0101 0010	52	R	114	0111 0010	72	r
19	0001 0011	13	[DC3]	51	0011 0011	33	3	83	0101 0011	53	S	115	0111 0011	73	s
20	0001 0100	14	[DC4]	52	0011 0100	34	4	84	0101 0100	54	T	116	0111 0100	74	t
21	0001 0101	15	[NAK]	53	0011 0101	35	5	85	0101 0101	55	U	117	0111 0101	75	u
22	0001 0110	16	[SYN]	54	0011 0110	36	6	86	0101 0110	56	V	118	0111 0110	76	v
23	0001 0111	17	[ETB]	55	0011 0111	37	7	87	0101 0111	57	W	119	0111 0111	77	w
24	0001 1000	18	[CAN]	56	0011 1000	38	8	88	0101 1000	58	X	120	0111 1000	78	x
25	0001 1001	19	[EM]	57	0011 1001	39	9	89	0101 1001	59	Y	121	0111 1001	79	y
26	0001 1010	1A	[SUB]	58	0011 1010	3A	:	90	0101 1010	5A	Z	122	0111 1010	7A	z
27	0001 1011	1B	[ESC]	59	0011 1011	3B	;	91	0101 1011	5B	[	123	0111 1011	7B	{
28	0001 1100	1C	[FS]	60	0011 1100	3C	<	92	0101 1100	5C	\	124	0111 1100	7C	|
29	0001 1101	1D	[GS]	61	0011 1101	3D	=	93	0101 1101	5D	]	125	0111 1101	7D	}
30	0001 1110	1E	[RS]	62	0011 1110	3E	>	94	0101 1110	5E	^	126	0111 1110	7E	~
31	0001 1111	1F	[US]	63	0011 1111	3F	?	95	0101 1111	5F	_	127	0111 1111	7F	[DEL]

Le pré-traitement consiste à garder en mémoire dans une liste la position de la dernière occurrence de chaque caractère distinct du motif (sauf le dernier, car il est utilisé pour comparer !)

Prenons un exemple de motif : 'CGGCTG' :

C : Le 'C' le plus à droite est en indice 3

G : Le 'G' le plus à droite est en indice 2

T : Le 'T' le plus à droite est en indice 4

Dans l'alphabet ASCII, on va attribuer 3,2 et 4 respectivement à 'C','G' et 'T'. Pour les 253 autres caractères, on va indiquer -1.

On aura donc :

Case 1	....	65	66	67	....	71	...	84	...	Case 256
-1	-1	-1-	-1	Ø –> 3	-1	1–>2	-1	4	-1	-1

Exercice 14.14
Effectuer le pré-traitement pour le motif 'DAB'

L'algorithme de pré-traitement

```
1  VARIABLE
2  motif : str
3  m : int # longueur du motif
4  DEBUT
5  Function PRE_ TRAITEMENT(motif)
6  |  doc = []
7  |  for i de 0 à 255 do
8  |  |  Ajouter -1 à d_oc
9  |  end
10 |  for i de 0 à m-2 do
11 |  |  doc[ord(motif[i])]← i
12 |  end
13 |  Retourner doc
14 end
```

Implémentation en Python

Exercice 14.15
Implémenter cet algorithme en Python. Vérifier avec les exemples précédents.

2.3.4 L'algorithme de Boyer Moore

L'algorithme

```
1  VARIABLE
2  motif : str
3  texte : str
4  m : int # longueur du motif
5  n :int # longueur du texte
6  i : int
7  DEBUT
8  Function BOYER_ MOORE(motif,texte)
9      liste = []
10     doc = PRE_ TRAITEMENT(motif)
11     i ← 0
12     Tant que i + m <= n Faire
13         j ← m - 1
14         Tant que j>= 0 and motif[j] == texte[i+j] Faire
15             j = j - 1
16         if j < 0 then
17             Ajouter i à liste
18             i ← i + m
19         else
20             i ← i + max(1,j - doc[ord(texte[i + j])])
21         end
22     Retourner liste
23 end
```

Implémentation Python

Exercice 14.16
Implémenter cet algorithme en Python. Vérifier avec les exemples précédents.

Remarque

Nous n'avons présenté qu'une version simplifiée de l'algorithme complet. L'algorithme complet de Boyer-Moore utilise une deuxième table de décalage, beaucoup plus difficile à calculer, qui permet de tenir compte des caractéristiques du motif dans le cas où celui-ci présente des similarités internes, ce qui permet d'effectuer des décalages plus importants, donc d'augmenter l'efficacité de la recherche. L'algorithme complet de Boyer-Moore présente des difficultés en termes de justification et de programmation effective qui dépassent le niveau attendu en NSI. C'est pourquoi nous ne l'évoquons pas ici. Le lecteur curieux pourra lire les pages 360–366 de l'ouvrage de Berstel, Beauquier et Chrétienne, disponible en ligne : ici