

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte. On notera :

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte. On notera :

ullet m le motif et l_m sa longueur

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte. On notera :

- ullet m le motif et l_m sa longueur
- t le texte et l_t sa longueur

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte.

On notera:

- ullet m le motif et l_m sa longueur
- ullet t le texte et l_t sa longueur

D'autre part, parfois le problème se réduira à déterminer si m est présent ou non dans t, ou encore on cherchera uniquement la première occurence.

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte.

On notera:

- m le motif et l_m sa longueur
- ullet t le texte et l_t sa longueur

D'autre part, parfois le problème se réduira à déterminer si m est présent ou non dans t, ou encore on cherchera uniquement la première occurence.

Exemple

La recherche du motif m=exe ($l_m = 3$) dans le texte t =un excellent exemple et un exercice extraordinaire ($l_t = 50$) doit produire la liste d'occurrences : [13; 27].

1. Position du problème

Définitions et notations

On s'intéresse au problème de la recherche d'une chaîne de caractères appelée motif dans une autre chaîne de caractères appelée texte. Plus précisement, on veut lister toutes les occurences (par leur position) du motif dans le texte. On notera :

- ullet m le motif et l_m sa longueur
- t le texte et l_t sa longueur

D'autre part, parfois le problème se réduira à déterminer si m est présent ou non dans t, ou encore on cherchera uniquement la première occurence.

Exemple

La recherche du motif m=exe ($l_m = 3$) dans le texte t =un excellent exemple et un exercice extraordinaire ($l_t = 50$) doit produire la liste d'occurrences : [13; 27].

 $\underset{\scriptscriptstyle{0}}{\text{un}}_{\sqcup} \text{excellent}_{\underset{\scriptscriptstyle{13}}{\sqsubseteq}} \text{exemple}_{\sqcup} \text{et}_{\sqcup} \text{un}_{\underset{\scriptscriptstyle{27}}{\sqsubseteq}} \text{exercice}_{\sqcup} \text{extraordinaire}$

2. Recherche naïve

Recherche naïve

Pour recherche si un motif m se trouve dans un texte t, on peut :

• parcourir chaque caractère de t jusqu'à celui d'indice ? inclus (indice de la dernière occurrence possible) :

indice dans le texte indice dans le motif

	,		
0		?:	$l_t - 1$
		0	l_m-1



Recherche naïve

Pour recherche si un motif m se trouve dans un texte t, on peut :

• parcourir chaque caractère de t jusqu'à celui d'indice l_t-l_m inclus (indice de la dernière occurrence possible) :

indice dans le texte indice dans le motif

, , ,	 -	
0	 $l_t - l_m$	$l_t - 1$
	0	l_m-1

Pour recherche si un motif m se trouve dans un texte t, on peut :

• parcourir chaque caractère de t jusqu'à celui d'indice l_t-l_m inclus (indice de la dernière occurrence possible) :

indice dans le texte 0 ...

	0	 $l_t - l_m$	$l_t - 1$
,		0	l_m-1

 $oldsymbol{2}$ si le caractère correspond au premier caractère du motif m, alors on avance dans le motif tant que les caractères coïncident.

Pour recherche si un motif m se trouve dans un texte t, on peut :

• parcourir chaque caractère de t jusqu'à celui d'indice $l_t - l_m$ inclus (indice de la dernière occurrence possible) : indice dans le texte $0 \ \ldots \ l_t - l_m \ l_t - 1$

indice dans le motif

0	 $l_t - l_m$	$l_t - 1$
	0	l_m-1

- $oldsymbol{2}$ si le caractère correspond au premier caractère du motif m, alors on avance dans le motif tant que les caractères coı̈ncident.
- ullet si on atteint la fin du motif, alors m se trouve dans t. Sinon on passe au caractère suivant de t.

Pour recherche si un motif m se trouve dans un texte t, on peut :

• parcourir chaque caractère de t jusqu'à celui d'indice $l_t - l_m$ inclus (indice de la dernière occurrence possible) : indice dans le texte $0 \ \ldots \ l_t - l_m \ l_t - 1$

indice dans le motif

Λ	 1 1	7 1
U	 $\iota_t - \iota_m$	$\iota_t - 1$
	0	l_m-1

- $oldsymbol{2}$ si le caractère correspond au premier caractère du motif m, alors on avance dans le motif tant que les caractères coı̈ncident.
- ullet si on atteint la fin du motif, alors m se trouve dans t. Sinon on passe au caractère suivant de t.

Exemple

Visualisation en ligne du fonctionnement de l'algorithme



Implémentation en OCaml

On renvoie la *liste* de toutes les occurrences : naive : string -> string -> int list

Implémentation en OCaml

```
On renvoie la liste de toutes les occurrences : naive : string -> string ->
 int list
   let naive motif texte =
      let lm = String.length motif
                                      in
      let lt = String.length texte in
      let occ = ref [] in
      for it=0 to lt-lm do
        let im = ref 0 in
        while (!im<lm && motif.[!im]=texte.[it+ !im]) do
          im := !im +1;
        done:
        if (!im=lm) then occ := it::!occ
10
        done;
11
      !occ;;
12
```



Implémentation en OCaml

Dans le cas où on teste simplement la présence, on peut provoquer une sortie anticipée de la boucle for à l'aide d'une exception.

Implémentation en OCaml

Dans le cas où on teste simplement la présence, on peut provoquer une sortie anticipée de la boucle for à l'aide d'une exception.

```
let present motif texte =
      let lm = String.length motif in
      let lt = String.length texte in
      try
      for it=0 to lt-lm do
        let im = ref 0 in
        while (!im<lm && motif.[!im]=texte.[it+ !im]) do
          im := !im +1:
        done:
        if (!im=lm) then raise Exit (* Exit est prédéfinie *)
10
      done;
11
      false
12
      with Exit -> true;;
13
```



Coût de la recherche simple

En notant l_m la longueur du motif et l_t celle de la chaine :



Coût de la recherche simple

En notant l_m la longueur du motif et l_t celle de la chaine :

ullet La boucle for est parcourue au plus l_t-l_m+1 fois

Coût de la recherche simple

En notant l_m la longueur du motif et l_t celle de la chaine :

- ullet La boucle for est parcourue au plus l_t-l_m+1 fois
- \bullet Pour chacun de ces parcours, la boucle while interne est parcourue au plus l_m fois

Coût de la recherche simple

En notant l_m la longueur du motif et l_t celle de la chaine :

- La boucle for est parcourue au plus $l_t l_m + 1$ fois
- ullet Pour chacun de ces parcours, la boucle while interne est parcourue au plus l_m fois

Au plus, l'algorithme effectue donc $l_m(l_t - l_m + 1)$ comparaisons.

Coût de la recherche simple

En notant l_m la longueur du motif et l_t celle de la chaine :

- La boucle for est parcourue au plus $l_t l_m + 1$ fois
- ullet Pour chacun de ces parcours, la boucle while interne est parcourue au plus l_m fois

Au plus, l'algorithme effectue donc $l_m(l_t - l_m + 1)$ comparaisons.

Exemple

Combien de comparaisons seront nécessaires si on recherche le motif bbbbbbbbb (neuf fois le caractère b suivi d'un a) dans une chaine contenant un million de b?



Accélération de la recherche

Supposons qu'on recherche le motif extra dans la chaine un excellent exemple et un exercice extraordinaire. La comparaison naïve ci-dessus commence par :



Accélération de la recherche

Supposons qu'on recherche le motif extra dans la chaine un excellent exemple et un exercice extraordinaire. La comparaison naïve ci-dessus commence par :

u	n		е	х	С	е	1	1	е	n	t
↓											
е	х	t	r	a							



Accélération de la recherche

Supposons qu'on recherche le motif extra dans la chaine un excellent exemple et un exercice extraordinaire. La comparaison naïve ci-dessus commence par :

 u
 n
 e
 x
 c
 e
 l
 l
 e
 n
 t

Deux idées vont permettre d'accélérer la recherche :

Accélération de la recherche

Supposons qu'on recherche le motif extra dans la chaine un excellent exemple et un exercice extraordinaire. La comparaison naïve ci-dessus commence par :

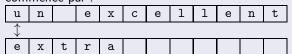
 u
 n
 e
 x
 c
 e
 l
 l
 e
 n
 t

Deux idées vont permettre d'accélérer la recherche :

• Commencer par la fin du motif.

Accélération de la recherche

Supposons qu'on recherche le motif extra dans la chaine un excellent exemple et un exercice extraordinaire. La comparaison naïve ci-dessus commence par :



Deux idées vont permettre d'accélérer la recherche :

- Commencer par la fin du motif.
- Prétraiter le motif de façon à éviter des comparaisons inutiles.



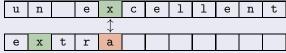
Accélération de la recherche

Dans l'exemple ci-dessus cela donne :

3. Algorithme de Boyer-Moore

Accélération de la recherche

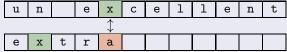
Dans l'exemple ci-dessus cela donne :



3. Algorithme de Boyer-Moore

Accélération de la recherche

Dans l'exemple ci-dessus cela donne :

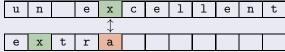


On peut directement décaler le motif de 3 emplacements car le dernier x du motif se trouve à 3 emplacements de la fin du motif.

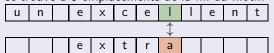
3. Algorithme de Boyer-Moore

Accélération de la recherche

Dans l'exemple ci-dessus cela donne :



On peut directement décaler le motif de 3 emplacements car le dernier x du motif se trouve à 3 emplacements de la fin du motif.



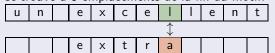
3. Algorithme de Boyer-Moore

Accélération de la recherche

Dans l'exemple ci-dessus cela donne :

	u	n		е	х	С	е	1	1	е	n	t
ľ	_											
	е	х	t	r	a							

On peut directement décaler le motif de 3 emplacements car le dernier x du motif se trouve à 3 emplacements de la fin du motif.



Cette fois, le 1 ne se trouve pas dans le motif, on peut donc décaler de la longueur du motif. Et la recherche s'arrête en ayant effectué seulement deux comparaisons.

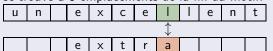


Accélération de la recherche

Dans l'exemple ci-dessus cela donne :

	u	n		е	х	С	е	1	1	е	n	t
ľ	_											
	е	х	t	r	a							

On peut directement décaler le motif de 3 emplacements car le dernier x du motif se trouve à 3 emplacements de la fin du motif.



Cette fois, le 1 ne se trouve pas dans le motif, on peut donc décaler de la longueur du motif. Et la recherche s'arrête en ayant effectué seulement deux comparaisons.

Visualisation en ligne

Visualisation en ligne du fonctionnement de l'algorithme accéléré



Algorithme de Boyer-Moore-Hoorspool

• La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :

3. Algorithme de Boyer-Moore

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la dernière occurrence de c et la fin du motif (l'avant dernière si c est le dernier caractère.)

3. Algorithme de Boyer-Moore

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la *dernière occurrence* de c et la fin du motif (l'avant dernière si c est le dernier caractère.)
 - Sinon la longueur du motif

3. Algorithme de Boyer-Moore

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la dernière occurrence de c et la fin du motif (l'avant dernière si c est le dernier caractère.)
 - Sinon la longueur du motif
- Ensuite on effectue une recherche en partant de la fin du motif en cas de non correspondance, on décale de la valeur fournie par la fonction de décalage.

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la *dernière occurrence* de c et la fin du motif (l'avant dernière si c est le dernier caractère.)
 - Sinon la longueur du motif
- Ensuite on effectue une recherche en partant de la fin du motif en cas de non correspondance, on décale de la valeur fournie par la fonction de décalage.

Exemple

3. Algorithme de Boyer-Moore

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la dernière occurrence de c et la fin du motif (l'avant dernière si c est le dernier caractère.)
 - Sinon la longueur du motif
- Ensuite on effectue une recherche en partant de la fin du motif en cas de non correspondance, on décale de la valeur fournie par la fonction de décalage.

Exemple

Construire la table de décalage du motif "toto"

3. Algorithme de Boyer-Moore

Algorithme de Boyer-Moore-Hoorspool

- La première phase consiste en un prétraitement du motif, afin de construire une fonction de décalage qui indique pour chaque caractère c :
 - Si c est dans le motif, Le nombre de caractères entre la dernière occurrence de c et la fin du motif (l'avant dernière si c est le dernier caractère.)
 - Sinon la longueur du motif
- Ensuite on effectue une recherche en partant de la fin du motif en cas de non correspondance, on décale de la valeur fournie par la fonction de décalage.

Exemple

- Construire la table de décalage du motif "toto"
- Simuler le fonction de l'algorithme de Boyer-Moore-Hoorspool pour recherche ce motif dans le chaine "zéro plus zéro = la tête à toto"

3. Algorithme de Boyer-Moore

Exercic<u>es</u>

- Ecrire en C, une fonction de signature int *cree_decalage(char *motif) qui renvoie la table de décalage d'un motif. On représentera un caractère par son code (donc un entier) et on suppose qu'on utilise la table ASCII standard qui contient 127 caractères.
- Simuler la recherche de abb dans le texte b...b (t fois la lettre b) avec l'algorithme de Boyer-Mooore. Donner le nombre de comparaisons effectué et comparer avec la recherche naïve.

Implémentation en C

Fonction qui renvoie true si motif est présent dans texte et false sinon.

```
bool appartient_bmh(char *motif, char *texte){
        int *dec = cree decalage(motif);
        int lt = strlen(texte);
        int lm = strlen(motif);
        int idx = 0:
        int im:
        while (idx < lt - lm + 1){
            im = lm - 1;
            while (im \geq 0 && texte[idx + im] == motif[im]){
                im = im - 1;
10
            if (im < 0){
11
                return true;}
12
            idx += dec[(int)texte[idx + lm - 1]];}
13
        return false;}
14
```

4. Algorithme de Rabin-Karp

Principe

L'idée de l'algorithme est d'utiliser une fonction de hachage h sur les chaines de caractères, et de comparer les hash du motif h(m) avec le hash du texte commençant à l'indice $\mathbf i$ et de longueur l_m pour $\mathbf i$ de 0 à l_t-l_m . On effectue la comparaison caractère par caractère seulement dans le cas où les deux hash sont égaux.