Recherche textuelle

Christophe Viroulaud

Terminale - NSI

Algo 27

Recherche textuelle

Approche naïve

Principe

Approche plu

Boyer-Moore

Recherche à l'enver

Décalages par saut

Prétraitement du motif

(simplifié - version

Recherche textuelle

Approche naïve

Principe

Approche plu efficace :

Boyer-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

lgorithme de Boyer-Moor simplifié - version

Complexit

La recherche textuelle est une fonctionnalité intégrée dans tous les logiciels de traitements de texte.



FIGURE 1 – Des applications multiples

- 4 bases nucléiques : Adénine, Cytosine, Guanine, Thymine,
- ► ADN humain : 3 milliards de bases répartis sur 23 paires de chromosomes.

Remarque

À titre de comparaison un roman compte environ 500000 caractères.

Recherche

Approche naïv

ficace : over-Moore

Boyer-Moore Recherche à l'envers

Décalages par sauts Prétraitement du motif

Algorithme de Boyer-Mooi (simplifié - version Horspool) Complexité

Comment effectuer une recherche textuelle efficace?

Recherche textuelle

Approche naïve

Principe

efficace : Boyer-Moore

Recherche à l'envers

Décalages par sauts

Algorithme de Boyer-Moore

Sommaire

Recherche textuelle

1. Approche naïve

- 1.1 Principe
- 1.2 Implémentation
- 2. Approche plus efficace: Boyer-Moore

Approche naïve

Implémentation

.

efficace :

D 1 1 1 1 1

D. ()

Decarages par sauts

Pretraitement du motif

Algorithme de Boyer-Moore simplifié - version Horspool)

Approche naïve - Principe

- b observer une **fenêtre** du texte,
- dans cette fenêtre, comparer chaque lettre du motif recherché au texte,
- décaler la fenêtre d'un cran dès qu'il n'y a pas de correspondance.

				3							
texte motif	а	С	g	а	t	С	С	a	t	g	a
motif	С	а	t								
	0	1	2								

Recherche textuelle

Approche naïve

Principe

Implémentation

Approche plus efficace :

Bover-Moore

Recherche à l'envers

Décalages par sauts

Algorithme de Boyer-Moore simplifié - version Horspool)

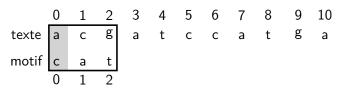


FIGURE 2 – Première comparaison : pas de correspondance

Décalage de la fenêtre

Recherche textuelle

Approche naïve

Principe

Implémentation

efficace : Bover-Moore

Recherche à l'envers

Décalages par saut

Prétraitement du n

gorithme de Boyer-Moore mplifié - version

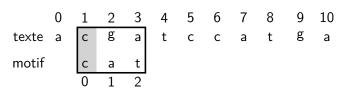


FIGURE 3 – Première comparaison : correspondance

Recherche textuelle

Principe

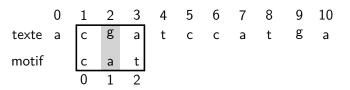


FIGURE 4 – Deuxième comparaison : pas de correspondance

Décalage de la fenêtre

Recherche textuelle

Principe

Sommaire

Recherche textuelle

Implémentation

- 1. Approche naïve
- 1.2 Implémentation

Activité 1:

- 1. Écrire la fonction recherche_naive(texte: str, motif: str) \rightarrow int qui renvoie la position du motif dans le texte ou -1 s'il n'est pas présent.
- 2. Estimer la complexité temporelle de cet algorithme dans le pire des cas : le motif n'est pas présent dans le texte.

Implémentation

```
def recherche_naive(texte: str, motif: str) -> int:
       11 11 11
       renvoie la position du motif dans le texte
       -1 s'il n'est pas présent
       11 11 11
       # dernière position = taille(texte) - taille(
6
     motif)
       for i in range(len(texte)-len(motif)+1):
           i = 0
           while j < len(motif) and
                   motif[j] == texte[i+j]:
               i += 1
           # correspondance sur toute la fenêtre
           if j == len(motif):
               return i
       return -1
```

Approche naïve

Implémentation

Approche plus efficace : Boyer-Moore Recherche à l'envers Décalages par sauts

> gorithme de Boyer-Moore mplifié - version rspool)

Imaginons le cas :

- On vérifie toute la fenêtre à chaque fois.
- À chaque **non correspondance** la fenêtre avance de 1.
- La complexité dépend de la taille du texte et de celle du motif.

Implémentation

Sommaire

Recherche textuelle

- 2. Approche plus efficace : Boyer-Moore

- 2.4 Algorithme de Boyer-Moore (simplifié version

Boyer-Moore

▶ 1970 : algorithme de Knuth-Morris-Pratt. O(T + M)

Recherche textuelle

Boyer-Moore

- 1970 : algorithme de Knuth-Morris-Pratt. O(T+M)
- ▶ 1977 : algorithme de Boyer-Moore.
 - ightharpoonup meilleur des cas : O(T/M)
 - ightharpoonup pire des cas : O(T+M)

Recherche textuelle

- ▶ 1970 : algorithme de Knuth-Morris-Pratt. O(T + M)
- ▶ 1977 : algorithme de Boyer-Moore.
 - ightharpoonup meilleur des cas : O(T/M)
 - pire des cas : O(T + M)
- ▶ 1980 : Horspool propose une version simplifiée de l'algorithme de Boyer-Moore. O(T)

La première idée de cet algorithme est de commencer la recherche **en partant de la fin du motif**.

 $\operatorname{Figure} 5$ – Première comparaison : pas de correspondance

Approche naïve

Implémentation

efficace :

Rover-Moore

Recherche à l'envers

Décalages par sau

Prétraitement du motif

gorithme de Boyer-Moore implifié - version

Remarque

Pour l'instant cette approche ne semble par apporter d'amélioration par rapport à l'algorithme précédent.

Recherche textuelle

Approche naïve

Principe

Approche plus efficace :

Recherche à l'envers

Décalages par saut

llgorithme de Boyer-Moo simplifié - version

Complexite

Sommaire

Recherche textuelle

1. Approche naïve

- 2. Approche plus efficace : Boyer-Moore
- 2.1 Recherche à l'envers
- 2.2 Décalages par sauts
- 2.3 Prétraitement du motif
- 2.4 Algorithme de Boyer-Moore (simplifié version Horspool)
- 2.5 Complexité

Approcne naive

Implémentation

Approche plus

Boyer-Moore

Recherche à l'envers

Décalages par sauts

Deáteoitament du m

Algorithme de Boyer-Moore simplifié - version

Décalage par sauts

Le motif ne contient pas la lettre ${\bf g}$ (la dernière lettre de la fenêtre).

FIGURE 6 – Comparaisons inutiles

Recherche textuelle

Approche naïve

Principe

efficace :

Recherche à

Décalages par sauts

Algorithme de Boyer-Moore

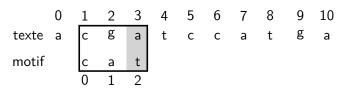


FIGURE 7 – Comparaison inutile

Recherche textuelle

Approche naïve

Principe

Implémentation

efficace : Boyer-Moore

Reclierche a l'enve

Décalages par sauts

Algorithme de Boyer-Moore (simplifié - version

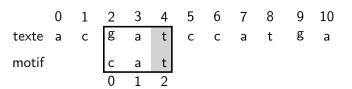


FIGURE 8 - Comparaison inutile

Recherche textuelle

Décalages par sauts

On peut donc directement décaler le motif à l'indice 3 du texte.

FIGURE 9 – Décalage par saut

Recherche textuelle

Approche naïve

1 1/

Approche

efficace : Boyer-Moore

Recherche à l'envi

Décalages par sauts

Algorithme de Boyer-Moore (simplifié - version

On n'observe pas de correspondance par contre la lettre **c** existe dans le motif. On va donc le décaler pour les faire coïncider.

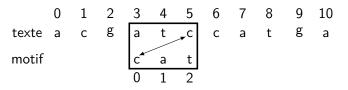


FIGURE 10 – Nouvelle situation

Recherche textuelle

Approche naïve

Principe

pproche pli

fficace : Boyer-Moore

Recherche a Lenvers

Décalages par sauts

Algorithme de Boyer-Moore

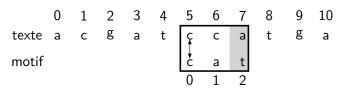


FIGURE 11 – Décalage par saut

Recherche textuelle

Approche naïve

Principe

Approche plu

Boyer-Moore

Décalages par sauts

Decarages par sauts

Algorithme de Boyer-Moore (simplifié - version

À retenir

On décale la position de recherche dans le texte en fonction de la dernière lettre de la fenêtre.

Recherche textuelle

Approche naïve

Principe

Implémentation

efficace : Boyer-Moore

Recherche à

Décalages par sauts

Decarages par sauts

Algorithme de Boyer-Moore (simplifié - version

Complexite

Sommaire

Recherche textuelle

1. Approche naïve

- 2. Approche plus efficace : Boyer-Moore
- 2.1 Recherche à l'envers
- 2.2 Décalages par sauts
- 2.3 Prétraitement du motif
- 2.4 Algorithme de Boyer-Moore (simplifié version Horspool)
- 2.5 Complexité

Approche naive

Implémentation

Approche plu

Boyer-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore

simplifié - version lorspool)

À retenir

Pour pouvoir décaler par saut, il faut connaître la dernière position de chaque lettre dans le motif. Le prétraitement consiste à calculer le décalage à appliquer pour amener chaque caractère du motif à la place du dernier caractère.

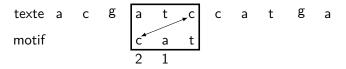


FIGURE 12 – Calculs des décalages

Approche naïve

Principe

Approche plus efficace :

Recherche à l'env

Décalages par sau

Prétraitement du motif

sigorithme de Boyer-ivi simplifié - version ·lorspool) ·omplevité

Remarque

On ne regarde pas la dernière position de la clé (la lettre t ici). Sinon la distance associée serait nulle et on resterait sur place après l'avoir lue dans le texte.



FIGURE 13 – Sauf la dernière lettre

Recherche textuelle

Approche naïve

Principe

Implémentation

Approche plus efficace :
Boyer-Moore

D/ 1

Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore (simplifié - version Horspool)

À retenir

Dans le cas de la répétition d'un caractère, on garde la distance la plus courte.

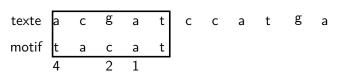


FIGURE 14 – Répétition dans le motif

Recherche textuelle

Approche naïve

Principe

Implémentation

efficace :

Recherche à l'en

Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore (simplifié - version Horspool)

Activité 2 : Écrire la fonction pretraitement_decalages(motif: str) \rightarrow dict qui associe chaque lettre du motif (sauf la dernière) à son décalage.

Recherche textuelle

Approche naïve

Principe

Implémentation

efficace : Boyer-Moore

recilercile a relivi

Decarages par sau

Prétraitement du motif

Igorithme de Boyer-Moore simplifié - version lorspool)

.1

return decalages

```
1
  def pretraitement decalages(motif: str) -> dict:
       renvoie le dictionnaire des décalages à
     appliquer
       pour chaque lettre du motif (sauf dernière)
       11 11 11
       decalages = dict()
6
       # on s'arrête à l'avant dernière lettre du motif
       for i in range(len(motif)-1):
           # la distance est mise à jour en cas de répé
     tition
           decalages[motif[i]] = len(motif)-1-i
```

Approche naïve

Implémentation

efficace :
Bover-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore (simplifié - version Horspool)

Sommaire

Recherche textuelle

- 2. Approche plus efficace : Boyer-Moore

- 2.4 Algorithme de Boyer-Moore (simplifié version Horspool)

Algorithme de Boyer-Moore (simplifié - version Horspool)

Algorithme de Boyer-Moore

L'algorithme de Boyer-Moore s'écrit alors :

```
Créer le tableau des décalages
Tant qu'on n'est pas à la fin du texte
Comparer le motif à la position du texte
Si le motif est présent
Renvoyer la position
Sinon
Décaler la fenêtre
Renvoyer -1 si le motif n'est pas présent
```

Code 1 – Algorithme de Boyer-Moore (version Horspool)

Recherche

Approche naïve

Implémentation

pproche plus ficace :

Recherche à l'envers

calages par sauts

Algorithme de Boyer-Moore (simplifié - version Horspool)

Activité 3:

- Écrire la fonction compare(texte: str, position: int, motif: str) → bool qui renvoie *True* si le motif est présent à la position i du texte.
- Écrire la fonction decalage_fenetre(decalages: dict, taille: int, lettre: str) → int qui renvoie le décalage à appliquer pour faire coïncider le motif à la dernière lettre de la fenêtre. Si la lettre n'est pas présente, la taille du motif est renvoyée.
- 3. Écrire alors la fonction boyer_moore(texte: str, motif: str) → int qui renvoie la position du motif dans le texte et -1 sinon.

Approche naïve
Principe

Approche plus

fficace :

Boyer-Moore

Recherche à l'envers Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore
(simplifié - version

Horspool) Complexité

```
def compare(texte: str, position: int, motif: str) -> bool:
 2
 3
       compare le morceau du texte
        (en partant de position + taille(motif))
 4
        avec le motif
 5
 6
 7
        Returns:
           bool: True si on a trouvé le motif
 q
        # position de la dernière lettre de la fenêtre
10
       en\_cours = position + len(motif) - 1
11
        # parcours de la fenêtre à l'envers
12
       for i in range(len(motif)-1, -1, -1):
13
           if not(texte[en_cours] == motif[i]):
14
              return False
15
           else:
16
17
              en cours -=1
        return True
18
```

Approche naïve

Implémentation

Approche pl

efficace : Boyer-Moore

Recherche à l'enver

Décalages par sauts

Algorithme de Boyer-Moore (simplifié - version

Horspool) Complexité

```
int:
       77 77 77
       renvoie la valeur du décalage à appliquer.
       si la lettre n'est pas dans le tableau
       c'est la taille du motif qui est appliqué
6
       Args:
           decalages (dict): dico des décalages
                                                                                       Algorithme de Boyer-Moore
                                                                                       (simplifié - version
           taille (int): taille du motif (= décalage max)
                                                                                       Horspool)
           lettre (str): dernière lettre de la fenêtre
       Returns:
.2
           int: décalage à appliquer
       77 77 77
       for cle, val in decalages.items():
           if cle == lettre:
.6
               return val
       # si la lettre n'est pas dans le dico (= le motif)
       return taille
```

def decalage_fenetre(decalages: dict, taille: int, lettre: str) ->

1

Recherche

textuelle

36 / 48

Recherche textuelle

```
def decalage fenetre2(decalages: dict, taille: int, lettre: str) ->
       int:
       # la méthode get renvoie une valeur par défaut si elle ne
       trouve pas la clé
       return decalages.get(lettre, taille)
5
   def decalage_fenetre3(decalages: dict, taille: int, lettre: str) ->
6
       int:
      try:
          res = decalages[lettre]
      except KeyError:
          res = taille
       return res
```

Approche naïve

rincipe molémentation

Approche plus efficace : Rover-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

Algorithme de Boyer-Moore

(simplifié - version Horspool)

```
def boyer_moore(texte: str, motif: str) -> int:
 1
 2
 3
        Returns:
           int: la position du motif dans le texte, -1 sinon.
 4
        77 77 77
 5
        decalages = pretraitement_decalages(motif)
 6
        i = 0
        while i \le len(texte) - len(motif):
 8
           # si on trouve le motif
 9
           if compare(texte, i, motif):
10
              return i
11
           else:
12
13
              # sinon on décale (en fonction de la dernière
        lettre de la fenêtre)
              decale = decalage fenetre(decalages,
14
                                     len(motif),
15
                                     texte[i+len(motif)-1])
16
              i += decale
17
        # si on sort de la boucle, on n'a rien trouvé
18
        return -1
19
```

Recherche

oproche naïve

Implémentation

pproche plu

Boyer-Moore

ócalagos nar sauts

Algorithme de Boyer-Moore (simplifié - version

Horspool) Complexité

Sommaire

Recherche textuelle

1. Approche naïve

- 2. Approche plus efficace : Boyer-Moore
- 2.1 Recherche à l'envers
- 2.2 Décalages par sauts
- 2.3 Prétraitement du motif
- 2.4 Algorithme de Boyer-Moore (simplifié version Horspool)
- 2.5 Complexité

Approche naive

Implémentation

efficace :

Boyer-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

implifié - version orspool)

Intuitivement l'algorithme semble plus rapide que la version naïve car il ne teste pas toutes les lettres du texte.

aaabaaabaaab

FIGURE 15 – Un cas représentatif

Approche naïve

Principe

Implémentation

efficace :
Bover-Moore

Recherche à l'envers

alages par sauts

Algorithme de Boyer-Moore

mplifié - version orspool)

FIGURE 16 - Algorithme na $\ddot{i}f$

Recherche textuelle

Approche naïve

Principe Implémentation

Implémentation

efficace :

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

gorithme de Boyer-Moore implifié - version orspool)

FIGURE 17 – Algorithme de Boyer-Moore

Recherche textuelle

Approche naïve

Principe Implémentation

Approche pl

efficace : Boyer-Moore

Décalages par caute

Prétraitement du n

Algorithme de Boyer-Moore (simplifié - version

Activité 4 : Compter le nombre d'itérations de la recherche avec l'algorithme naïf puis celui de Boyer-Moore.

Recherche textuelle

Approche naïve

Principe

Approche pl

Boyer-Moore

Recherche à l'envers

Décalages par sauts

Prétraitement du motif

(Igorithme de Boyer-Moor simplifié - version

Correction

Recherche textuelle

► Algorithme naïf : 10 décalages,

Algorithme de Boyer-Moore : 3 décalages.

Approche naïve

Principe

Implémentation

efficace :

Recherche à l'envers

Décalages par sauts

Prétraitement du r

Algorithme de Boyer-Moore simplifié - version

Remarques

- ▶ Dans le meilleur des cas, la complexité temporelle de l'algorithme est O(T/M) où T est la taille du texte et M celle du motif.
- ▶ Plus le motif est long plus l'algorithme est rapide.

Recherche textuelle

Approche naïve

Implémentation

Approche plus efficace :

Boyer-Moore
Recherche à l'envers

Décalages par sauts

Algorithme de Boyer-Moore

Recherche textuelle

Approche naïve

Principe

Approche pl

efficace : Boyer-Moore

Recherche à l'envers

Décalages par sauts

Algorithme de Boyer-Moore

nplifié - version rspool)

Complexité

Le prétraitement a un coût (temporel et spatial) mais qui est grandement compensé.

Cas critique

 ${\rm Figure}\ 18-Cas\ critique$

Recherche textuelle

Approche naïve

Principe Implémentation

Approche plus efficace :

Recherche à l'envers

Décalages par sauts

Algorithme de Boyer-Moor

spool)

Code complet

Les programmes du cours sont téléchargeables ici.

Recherche textuelle