



Télécharger sur <https://lycee.editions-bordas.fr>



Cours en podcast PAGE Flashable

4. Les moteurs de recherche

Les navigateurs web proposent à leurs utilisateurs d'accéder à des pages sans connaître leur URL et même sans en connaître l'existence à l'avance ! Pour cela, il est fait appel à des moteurs de recherche. Les moteurs de recherche parcourent en permanence le web, répertorient les pages et les classent en fonction de leur popularité.

Indexation du web

Les sociétés qui proposent des moteurs de recherche doivent disposer de serveurs très puissants et de capacités mémoire très importantes, car l'indexation demande d'enregistrer une copie de toutes les pages web accessibles au niveau mondial.

☞ **Indexation :** traitement qui consiste à analyser des pages pour y détecter des mots clés utilisés couramment dans les demandes des internautes, puis à fabriquer un index permettant de retrouver rapidement ces pages à partir des mots clés.

Tout ceci est calculé à l'avance, ce qui permet aux moteurs de recherche de répondre en un temps très court quand un nouvel utilisateur effectue une nouvelle demande. La demande est en fait rarement nouvelle : d'autres l'avaient probablement déjà effectuée avant.

Calcul de la popularité

La principale difficulté des moteurs de recherche consiste à classer, de la manière la plus pertinente possible l'ensemble des pages contenant les mots clés demandés, pour choisir quelles pages présenter en premier.

● C'est essentiellement sur ces algorithmes de classement que les moteurs de recherche se font une concurrence acharnée et cachent leurs secrets de fonctionnement. En 1998, les informaticiens américains Larry Page et Sergey Brin proposent l'algorithme *PageRank* qui a conduit à la création du moteur de recherche Google et permis de proposer une définition à l'idée de « popularité » d'une page.

● L'algorithme *PageRank* (de l'anglais *rank* : score) repose sur le principe de calculer la popularité d'une page à partir de la popularité des pages qui la citent. Cela évite de chercher à comprendre ce qui est écrit dans une page, ce qui serait difficile pour une machine. L'idée est de faire confiance aux auteurs des pages web : en citant une autre page, ils lui attribuent de la confiance. Cette confiance a de la valeur si elle émane d'un site lui-même très populaire.

☞ **Popularité :** plus une page est citée, plus sa popularité est grande, surtout si elle est citée par des pages ayant elles-mêmes une bonne popularité.

Ce calcul est en fait très complexe car il oblige à calculer en même temps la popularité de toutes les pages du web qui dépendent les unes des autres, ce qui aboutit à une gigantesque équation avec un nombre d'inconnues égal au nombre de pages du web !

eureka!

Dans la pratique, on simule un internaute qui suivrait des liens au hasard. La probabilité d'atteindre ainsi une page donne une approximation de sa popularité.



Identifiez la page la plus populaire et programmez p63

er sur <https://lycee.editions-bordas.fr>

Activité 10.

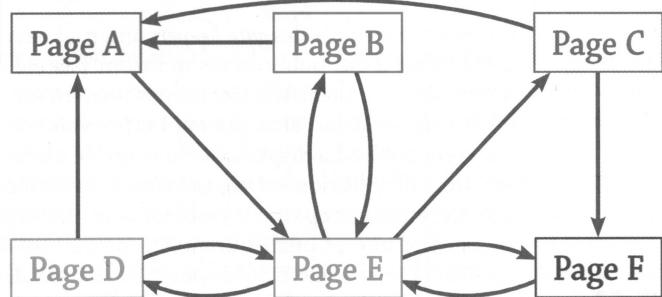
 Télécharger
Chap3_Act10.pdf

Algorithme du PageRank

Un moteur de recherche a enregistré six pages web. On a représenté par des arcs les hyperliens entre les différentes pages sur la figure voir > Doc. 1.

- Les pages A et F ont un hyperlien sortant.
- Les pages B, C et D ont deux hyperliens sortants.
- La page E a quatre hyperliens sortants.

Doc. 1 Pages et hyperliens



Afin de quantifier la popularité d'une page web, l'algorithme du PageRank calcule un score qui est proportionnel au nombre de fois qu'un internaute visite la page quand il clique de manière aléatoire sur les hyperliens.

Partie 1. Le jeu du PageRank

 Pour choisir au hasard entre deux destinations possibles, on peut tirer à pile ou face.

Pour choisir entre quatre destinations possibles, on peut faire deux lancers de pièce.

On se propose de jouer en classe au jeu dont voici les règles.

1. Tout d'abord, mettre un pion sur une page de départ au choix.
 2. Choisir au hasard une destination parmi les hyperliens sortants de cette page.
 3. Recommencer 50 fois à partir de la page atteinte, en notant le nombre de passage par chaque page.
- On obtient par exemple A : 8 ; B : 4 ; C : 6 ; D : 4 ; E : 21 ; F : 7.
4. a. Mettre en commun les résultats et calculer le score d'une page en faisant pour chaque page la somme des nombres de visites obtenus par tous les élèves de la classe.
 - b. Vérifier que les pourcentages obtenus par la classe sont proches de :
A : 15 % ; B : 10 % ; C : 10 % ; D : 10 % ; E : 40 % ; F : 15 %.

Partie 2. Analyse des résultats

- a. Quelle signification peut-on donner aux pourcentages calculés ?
- b. Quelle est la page qui a la popularité la plus grande ?
- c. Comment expliquer ce succès ?

En répétant un grand nombre de fois le jeu, on remarque que la popularité de la page F (15 %) est égale à la popularité de la page E divisée par quatre (10 %) plus la popularité de la page C divisée par deux (5 %).

2. a. Pourquoi faut-il diviser par quatre la popularité de E ?
- b. Pourquoi faut-il diviser par 2 la popularité de C ?
3. En raisonnant comme pour la question précédente, à combien la popularité de la page E doit-elle être égale ?

Pour aller plus loin



L'activité proposée ici nécessite de manipuler des tableaux pour représenter les pages et les liens entre les pages. Ces informations sont données pour que l'activité soit réalisable. Les algorithmes et la programmation des tableaux sont au programme de la spécialité « Numérique et Science Informatique » en classe de première.

→ NSI

Activité 11. Programmation du PageRank

Cette activité permet d'expérimenter la mise en œuvre de l'algorithme PageRank dans une version très simplifiée. Les six pages sont représentées chacune par leur nom de « A » à « F » qui sont enregistrés dans le tableau `nom`. Les hyperliens sont représentés aussi dans un tableau `hyperliens`.


Télécharger
Chap3_PageRank.py

```
from random import choice
nom = ["A", "B", "C", "D", "E", "F"]
hyperliens = [[4],[0,4],[0,5],[0,4],[1,2,3,5],[4]]
nbEtapes = 10 # on répète 10 fois
nbVisites = [0,0,0,0,0,0]
page = 0
for i in range(nbEtapes):
    page = choice(hyperliens[page])
    print("Page actuelle : " + nom[page])
```

1. Télécharger le programme mis à disposition. L'exécuter plusieurs fois et noter les résultats.
2. Expliquer ce qu'affiche ce programme.
3. Modifier le programme pour qu'il affiche le nombre de fois où l'on passe par la page E.
4. Modifier le programme pour qu'il calcule le score (un pourcentage) de chaque page en utilisant la liste `nbVisites`. Essayer plusieurs valeurs pour `nbEtapes`.
5. a. Vérifier que si l'on remplace le lien de la page A vers la page E par un lien de la page A vers la page C, les popularités de A et C augmentent.
b. Expliquer ce phénomène.

