Université Paris 13

IUT de Villetaneuse

Classement des pages web par Google

Ivan Lin – Alexandre Oudni – Kevin Wong

DUT Informatique

Projet MA

2017-2018

Table des matières

1 Introduction 1

2 Développement mathématiques

2.1

2.2

3 Algorithme mathématique

3.1 Pseudo-langage. . . . . . . . . . .

4 Différents outils informatique utilisé

4.1 Choix des technologies

4.2Explication du développement et du fonctionnement de l’interface graphique

5 Résultat obtenus

5.1 Tableau des résultats

6 Résultat obtenus

6.1 Principaux résultat

6.2 Objectifs réalisé

6.3 Perspectives

6.4 Difficultés rencontrées

7 Bibliographie

Annexes

1 Introduction

1.1Présentation

Dans le cadre des Projets S4, nous avons choisi le sujet du classement des pages web par Google. En 1999 Sergey Brin et Larry Page ont mis au point un algorithme mathématique appelé PageRank qui a pour but de déterminer le degré d’importance d’une page Web. Il se démarque des autres moteurs de recherche car il propose un classement des pages en fonction de leur indice de popularité. Celui-ci ne dépend pas du nombre de consultations de la page mais du nombre de liens qui pointent sur cette page à partir d’autres pages Web. En Novembre 2016 le nombre de pages indexées par Google est de 130 mille milliards de pages web.

Le projet a pour but de représenter la structure du Web par une matrice C ∈ Mn,n(R+) , pour cela nous devons trouver une méthode, un algorithme qui va calculer la pertinence de chaque page. On peut représenter la pertinence par un nombre ou un score positif avec la convention que plus le score est grand plus la page est importante.

2 Développement mathématiques

L’idée est de travailler sur le fait qu’il y a des liens reliant les pages les unes aux autres. La première chose à faire :

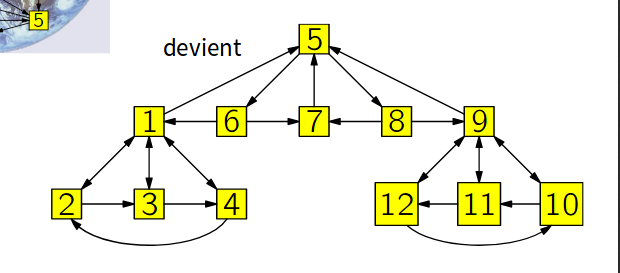
- Numéroter toutes les pages : p1, p2, ….. ,pj

Comme le but est de définir un critère de pertinence de la page, il faudra distinguer « p1 contient un lien qui pointe sur p2 » et « p2 contient un lien qui pointe sur p1 ».

- Si la page pj contient un lien vers la page pi on matérialise cela par une flèche pj → pi .

Considérons un réseau à 12 pages modélisé par le graphe suivant :





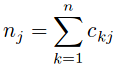
On peut créer la matrice C associée à ce graphe telle que Cij = 1 si la page j pointe sur la page i et 0 sinon. De plus, une page ne peut pas avoir de lien qui pointe sur elle-même, donc Cii = 0.

C =

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

i en ligne et j en colonne.

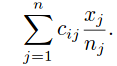
Chaque page i dispose d’un indice xi de façon à pouvoir classer l’ensemble des pages web par score décroissant. Nous devons donc calculer l’ensemble des indices xi en partant du principe qu’un lien de la page j pointant sur la page i contribue au score de cette dernière avec une pondération par xj (une page ayant un indice élevé a plus de poids qu’une page avec un mauvais



indice) et par le nombre total de liens nj présent sur la page

(une page avec beaucoup de lien a moins d’influence).

On a xi =



|  |
| --- |
| x1= x2/2 + x3/2 + x4/2 + x6/2 |
| x2= x1/4 + x4/2 |
| x3= x1/4 + x2/2 |
| x4= x1/4 + x3/2 |
| x5= x1/4 + x7/1 +x9/4 |
| x6= x5/2 |
| x7= x6/2 + x8/2 |
| x8= x5/2 |
| x9= x8/2 + x10/2 +x11/2 + x12/2 |
| x10= x9/4 + x12/2 |
| x11= x9/4 + x10/2 |
| x12= x9/4 + x11/2 |

Prenons le vecteur X =

|  |
| --- |
| x1 |
| x2 |
| x3 |
| x4 |
| x5 |
| x6 |
| x7 |
| x8 |
| x9 |
| x10 |
| x11 |
| x12 |

Prenons une matrice A =

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1/2 | 1/2 | 1/2 | 0 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/4 | 0 | 0 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/4 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/4 | 0 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1/4 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | 1/2 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1/2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | 1/2 | 1/2 | 1/2 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/4 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/4 | 1/2 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/4 | 0 | 1/2 | 0 |

Définition : v est un vecteur propre de A si il existe λ ∈ ℝ tel que Av = λv.

λ est appelée valeur propre associée à v.

Le problème du classement des pages du Web se trouve ramené à la recherche d’un vecteur associé à la valeur λ=1 de la matrice A.

AX = X

A – X = 0

(A – I)X = 0

Cependant, il est possible qu’un réseau contiennent des pages qui n’ont pas de liens (ce qui n’est pas le cas dans l’exemple que nous traitons) ; on se retrouverai donc avec une matrice A contenant des lignes pleines de 0. Dans ce cas , nous rajoutons des liens artificiels pondérés par 1/n qui tend vers 0 et on considère la matrice P telle que pij = (1/nj)Cij + (1/n)dj où d est tel que dj=1 si nj = 0 (si il n’y a pas de liens) et 0 sinon.

La matrice P est la transposée d’une matrice stochastique, pour eT = (1,…..,1) ∈ ℝn, nous avons eT\*P = eT, ce qui signifie que P admet bien la valeur propre 1.

Pour être sûr que la valeur propre est simple,

on pose A = αP + (1−α)(1/n)eeT où 0<α<1

4.Nous avons choisi de développer le projet sous Python3, avec comme éditeur de texte Sublime Text.

5.