

ALGÈBRE LINÉAIRE

COURS DU JEUDI 19 DÉCEMBRE

Jérôme Scherer



LA MATRICE DE GOOGLE

Soient P_i les (25 milliards de) pages du web. On pose

- ❶ l_j le nombre de liens d'une page P_j .
- ❷ B_i l'ensemble des pages ayant un lien vers P_i .

LA MATRICE DE GOOGLE

Soient P_i les (25 milliards de) pages du web. On pose

- ❶ l_j le nombre de liens d'une page P_j .
- ❷ B_i l'ensemble des pages ayant un lien vers P_i .

DÉFINITION

Le **rang** de la page P_i vaut
$$r(P_i) = \sum_{B_i} \frac{r(P_j)}{l_j}.$$

LA MATRICE DE GOOGLE

Soient P_i les (25 milliards de) pages du web. On pose

- ❶ l_j le nombre de liens d'une page P_j .
- ❷ B_i l'ensemble des pages ayant un lien vers P_i .

DÉFINITION

Le **rang** de la page P_i vaut
$$r(P_i) = \sum_{B_i} \frac{r(P_j)}{l_j}.$$

Il faut connaître tous les $r(P_j)$ pour calculer $r(P_i)$?

SERGEY BRIN ET LARRY PAGE

C'est le point de départ de l'idée géniale de deux doctorants de Stanford, vers 1995.



UN SYSTÈME D'ÉQUATIONS

Soit H la matrice dont les coefficients

$$H_{ij} = \begin{cases} \frac{1}{l_j} & \text{si } P_j \in B_i \\ 0 & \text{sinon} \end{cases}$$

UN SYSTÈME D'ÉQUATIONS

Soit H la matrice dont les coefficients

$$H_{ij} = \begin{cases} \frac{1}{l_j} & \text{si } P_j \in B_i \\ 0 & \text{sinon} \end{cases}$$

Soit $\vec{r}$ le vecteur “PageRank” dont les coefficients sont $r(P_i)$. Alors

$$\boxed{\vec{r} = H\vec{r}}$$

UN SYSTÈME D'ÉQUATIONS

Soit H la matrice dont les coefficients

$$H_{ij} = \begin{cases} \frac{1}{l_j} & \text{si } P_j \in B_i \\ 0 & \text{sinon} \end{cases}$$

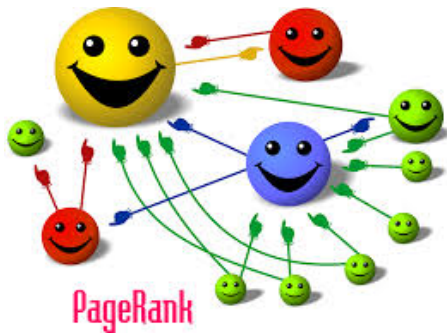
Soit $\vec{r}$ le vecteur “PageRank” dont les coefficients sont $r(P_i)$. Alors

$$\boxed{\vec{r} = H\vec{r}}$$

OBSERVATION

Le vecteur PageRank est un vecteur propre pour la valeur propre 1.

REPRÉSENTATION



PROBLÈMES

Pages “dead end” et sous-réseaux.

CORRECTIONS

Pour régler leur compte aux pages sans issues, on remplace les colonnes de zéros par des colonnes de $1/n$ de sorte que la somme des coefficients de chaque colonne vaut 1. Cette matrice est appelée S .

CORRECTIONS

Pour régler leur compte aux pages sans issues, on remplace les colonnes de zéros par des colonnes de $1/n$ de sorte que la somme des coefficients de chaque colonne vaut 1. Cette matrice est appelée S .

CONSÉQUENCE

Le nombre 1 est valeur propre de la matrice S .

CORRECTIONS

Pour régler leur compte aux pages sans issues, on remplace les colonnes de zéros par des colonnes de $1/n$ de sorte que la somme des coefficients de chaque colonne vaut 1. Cette matrice est appelée S .

CONSÉQUENCE

Le nombre 1 est valeur propre de la matrice S .

Pour régler le problème des sous-réseaux, on choisit un coefficient c (probablement $\approx 0,85$). Soit E la matrice dont tous les coefficients valent 1.

CORRECTIONS

Pour régler leur compte aux pages sans issues, on remplace les colonnes de zéros par des colonnes de $1/n$ de sorte que la somme des coefficients de chaque colonne vaut 1. Cette matrice est appelée S .

CONSÉQUENCE

Le nombre 1 est valeur propre de la matrice S .

Pour régler le problème des sous-réseaux, on choisit un coefficient c (probablement $\approx 0,85$). Soit E la matrice dont tous les coefficients valent 1.

$$G = cS + \frac{1-c}{n}E$$

LE SURFEUR DU WEB



Un surfeur du web sur la page P_i a une probabilité c de cliquer sur un lien de cette page, et une probabilité $1 - c$ de recommencer sa recherche d'informations sur une page choisie aléatoirement.

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.
- ❷ 1 est la plus grande valeur propre parmi toutes les valeurs propres, réelles et complexes.

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.
- ❷ 1 est la plus grande valeur propre parmi toutes les valeurs propres, réelles et complexes.
- ❸ La dimension de E_1 vaut 1.

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.
- ❷ 1 est la plus grande valeur propre parmi toutes les valeurs propres, réelles et complexes.
- ❸ La dimension de E_1 vaut 1.
- ❹ La deuxième valeur propre de G est de grandeur environ c .

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.
- ❷ 1 est la plus grande valeur propre parmi toutes les valeurs propres, réelles et complexes.
- ❸ La dimension de E_1 vaut 1.
- ❹ La deuxième valeur propre de G est de grandeur environ c .
- ❺ La matrice G est diagonalisable.

MATRICES STOCHASTIQUES, THÉORÈME DE PERRON

La matrice Google est stochastique et strictement positive.

- ❶ 1 est valeur propre.
- ❷ 1 est la plus grande valeur propre parmi toutes les valeurs propres, réelles et complexes.
- ❸ La dimension de E_1 vaut 1.
- ❹ La deuxième valeur propre de G est de grandeur environ c .
- ❺ La matrice G est diagonalisable.

CONSÉQUENCE

Le vecteur PageRank est bien défini ! La somme des $r(P_i)$ vaut 1.

CALCUL DU PAGERANK

Il est impossible de résoudre un système de 25 milliards d'équations à 25 milliards d'inconnues.

CALCUL DU PAGERANK

Il est impossible de résoudre un système de 25 milliards d'équations à 25 milliards d'inconnues.

On calcule une approximation par une méthode inductive. On choisit un vecteur $\vec{r}_0$ et on calcule

$$\boxed{\vec{r}_1 = G \vec{r}_0}$$

CALCUL DU PAGERANK

Il est impossible de résoudre un système de 25 milliards d'équations à 25 milliards d'inconnues.

On calcule une approximation par une méthode inductive. On choisit un vecteur $\vec{r}_0$ et on calcule

$$\boxed{\vec{r}_1 = G \vec{r}_0}$$

et on itère. Comme toutes les valeurs propres autres que 1 sont plus petites que 1 strictement, la partie de $\vec{r}_n$ relative à celles-ci tend vers zéro.

CALCUL DU PAGERANK

Il est impossible de résoudre un système de 25 milliards d'équations à 25 milliards d'inconnues.

On calcule une approximation par une méthode inductive. On choisit un vecteur $\vec{r}_0$ et on calcule

$$\boxed{\vec{r}_1 = G \vec{r}_0}$$

et on itère. Comme toutes les valeurs propres autres que 1 sont plus petites que 1 strictement, la partie de $\vec{r}_n$ relative à celles-ci tend vers zéro.

CONSÉQUENCE

Le vecteur PageRank $\vec{r} = \lim \vec{r}_n$.

JOYEUX NOËL !

