11020 BOES Rémi

Le PageRank

Curieux de comprendre le classement des différents sites web lors d'une recherche internet, je me suis intéressé au modèle du PageRank utilisé par Google.
J'ai donc décidé d'étudié la modélisation du PageRank ainsi que la mise en place d'algorithmes de résolution de celui-ci.
Cette étude s'inscrit dans le thème "Enjeux societaux" puisque le classement des sites web est un défi du 21ème siècle. En effet, être le mieux classé permet aux sites d'être les plus visible et ainsi pouvoir pour les sites marchands par exemple obtenir plus de clients.

Positionnement thématique (ETAPE 1)

MATHEMATIQUES (Mathématiques Appliquées), INFORMATIQUE (Informatique pratique).

Mots-clés (ETAPE 1)

Mots-Clés (en français)	**Mots-Clés** (en anglais)
PageRank	*PageRank*
Résolution	*Resolution*
Modélisation	*Modeling*
Comparaison	*Comparison*
Convergence	*Convergence*

Bibliographie commentée

Avec l'explosion d'internet, le partage d'informations nécessite la mise en place de moteurs de recherche afin de trouver les pages les plus pertinentes lors de nos recherches. L'innovation qui a aidé à l'explosion de Google fut la mise en place d'une modélisation mathématique : le PageRank.

L'idée derrière cette modélisation est d'imaginer le déplacement d'un surfer aléatoire se déplaçant de page en page en utilisant les liens de celles-ci. Au bout d'un certain temps, sa position la plus probable représentera le site qui sera intuitivement le mieux classé et ainsi de suite avec les autres sites, cette probabilité finale a été nommée le PageRank.

La modélisation mathématique consiste à concevoir le web comme un graphe où chaque page représente un sommet, et un arc entre 2 sommets correspond à un lien d'une page vers une autre. De cette manière on construit une matrice stochastique représentant les liens du web.
Pour palier certains problèmes celle-ci est légèrement modifiée pour devenir notre matrice de transition régissant les différentes marches de notre surfer entre les sites. La valeur que l'on nomme PageRank est ici la valeur stationnaire u tel que si A est notre matrice de transition : u=Au. [1]

Cependant, le nombre de pages web étant très grand (de l'ordre de 10^10) la résolution exacte du PageRank serait bien trop longue par des ordinateurs classiques, d'autant plus que celle-ci se doit d'être tenue à jour au maximum. Pour pallier ce problème de nombreuses méthodes de résolutions alternatives existent.

Parmi celles-ci se trouvent les méthodes dites " itératives " qui utilisent l'itération de suite qui converge vers notre PageRank et ceci de manière certaine d'après le Théorème de Perron-Frobenius. [2]

La premier algorithme étudié est celui dit de " la puissance " consistant à itérer une suite (Un) tel que Un+1=AUn qui converge vers notre valeur stationnaire recherchée u.
La résolution de notre PageRank est équivalente à résoudre un système de la forme Ax=b, celà nous permet alors d'utiliser des méthodes de résolution de systèmes linéaires.
Les méthodes dites de " Gauss-Seidel " et de " Jacobi " permettent de résoudre le système en utilisant la décomposition de A en matrices : diagonale, triangulaire supérieure et inférieure strictement.
Ces méthodes peuvent ensuite être " relaxées " en ajoutant un facteur d'amortissement w permettant d'accélérer la convergence à condition qu'il soit bien choisi.[3][4]

De manière plus général, il existe aussi des méthodes dites d'extrapolation qui permettent d'accélérer la convergence de suites et peuvent ici être utilisées. Elle se base sur l'estimation - pour n choisi - de Un comme combinaison linéaire des 2 premiers vecteurs propres de A pour la méthode d'extrapolation d' Aitken et des 3 premiers pour la méthode d'extrapolation quadratique. [5]

Problématique retenue

La modélisation mathématique du PageRank demande de mettre en place des stratégies de résolution de systèmes linéaires et l'utilisation de méthodes itératives permet cette résolution. Quel algorithme est le plus adapté pour résoudre un tel système ?

Objectifs du TIPE

Je me propose de :

- Mettre en place la modélisation du PageRank sous forme matricielle
- Implémenter les différentes méthodes itératives de résolution du PageRank et mettre en avant leur fonctionnement et leurs spécificités.
- Comparer à l'aide de graphes de différentes tailles la complexité spatiale et temporelle des différentes méthodes.

Références bibliographiques (ETAPE 1)

[1] Michael Eisermann, Université Joseph-Fourier de Grenoble, 2009 : Comment fonctionne Google ? : *https://www-fourier.ujf-grenoble.fr/~faure/enseignement/systemes_dynamiques/Documents_annexes/08_Eisermann_google_rank.pdf*

[2] Fabien Priziac, Université Aix-Marseille : Matrices stochastiques et théorèmes de Perron-Frobenius (cours licence) : *http://www.i2m.univ-amu.fr/~priziac.f/cma-CM15.pdf*

[3] Thierry Gallouet, Université Aix-Marseille : Méthodes itératives (cours licence) : *https://www.i2m.univ-amu.fr/perso/thierry.gallouet/licence.d/anum.d/anum-c4.pdf.*

[4] Jean Deteix, Université Laval, 2017 : Notes de cours Préconditionnement (cours) :

https://www2.mat.ulaval.ca/fileadmin/Cours/MAT-17992/notes_precond.pdf

[5] C. Manning G. Golub S. Kamvar T. Haveliwala, Université Stanford, 2003 : Extrapolation methods for accelerating pagerank computations :
http://infolab.stanford.edu/~taherh/papers/extrapolation.pdf

DOT

[1] *Novembre 2020 : Modélisation de la marche aléatoire*

[2] *Décembre 2020 : Etude de la construction de la matrice de transition de google et étude théorique du modèle*

[3] *Janvier 2021 : Mise en place de l'algorithme de la puissance*

[4] *Février 2021 : Création d'un programme de génération de matrice web*

[5] *Mars 2021 : Implémentation des différentes méthodes de résolution du PageRank (Jacobi, Gauss-Seidel, ...)*

[6] *Avril 2021 : Comparaison des méthodes et tracés des différents graphiques utiles à la présentation. Utilisation de matrices "réelles" plutôt que celles générées par mon programme*

[7] *Mai 2021 : Fin des tracés et des comparaisons des différentes méthodes*