Compte-rendu du projet Ranking & Recommandation

MACHE Ethan et JOLY Théo

Mai 2025

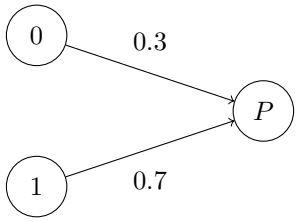
1 Explication du sujet 8

Dans le modèle de PageRank classique, lorsqu'une personne atteint une page sans lien sortant, on suppose qu'il saute ensuite de manière uniforme vers n'importe quelle autre page du Web. Cette hypothèse, bien que théoriquement cohérente, ne reflète pas toujours le comportement réel : en pratique, l'utilisateur appuie plutôt sur le bouton précédent de son navigateur pour retrouver la page qu'il venait de visiter. Le sujet 8 propose donc une modification du graphe Web afin de mémoriser la provenance de chaque visite sur une page sans lien sortant, et de simuler ainsi fidèlement ce retour en arrière.

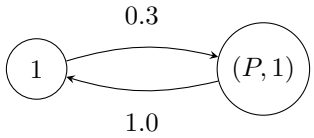
2 Traitements des sommets sans arcs sortants

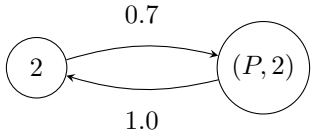
Pour réaliser cette modification, il faut transformer tous les sommets dépourvus d'arcs sortants en sommets pointant vers leur parent immédiat. Autrement dit, on crée pour chaque page un arc dédié vers la page précédente : lorsqu'une personne utilise la fonction précédent, il revient toujours vers la page qu'il venait de visiter, et non vers une autre page qui pointerait vers elle.

On aura donc :



Qui deviendra :



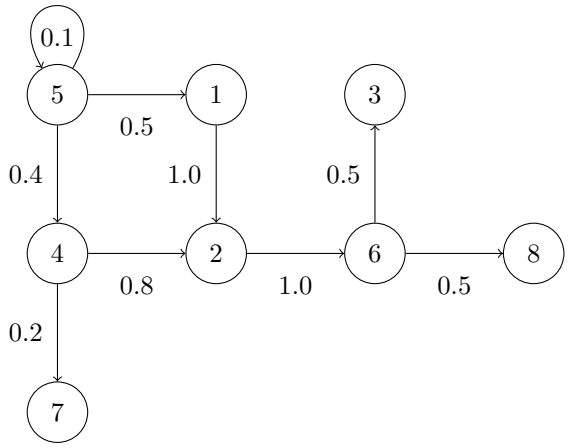


Sur ces graphiques, on constate que la vitesse de convergence diminue nettement à mesure que le nombre de sommets avec un degré sortant nul (sommets sans arc sortant) augmente. En effet, lorsque le pourcentage de ces sommets croît, la méthode Backspace met de plus en plus de temps à converger : pour $\alpha = 0{,}9$, elle passe de 20 itérations dans le cas sans sommets avec un degré sortant nul à 121 itérations lorsque leur proportion est maximale (soit plus de 10 fois plus lent), tandis que le PageRank classique reste stable, autour de 15 itérations.

Cette différence s'explique par le traitement des sommets de degré sortant nul, qui sont répartis en plusieurs nœuds reliés par des arcs de probabilité unitaire. Or, distribuer une probabilité de 1 sur l'ensemble des voisins est plus long avant d'atteindre l'équilibre. C'est pourquoi il est recommandé d'initialiser la distribution stationnaire en répartissant uniformément la probabilité sur tous les états, plutôt que d'attribuer d'emblée une probabilité de 1 à un unique sommet et 0 aux autres.

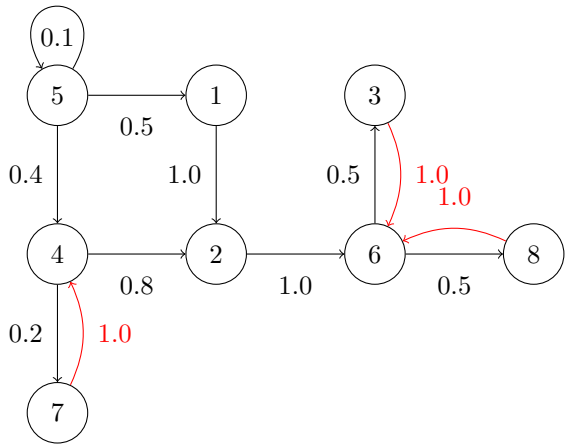
3 Exemple et interprétation des résultats

Graphe avant transformation



Quand on fait tourner PageRank classique sur le graphe on obtient cela :
Pour $\alpha = 0.85$:
$x = [0.085213, 0.184886, 0.149705, 0.079807, 0.063591, 0.215339, 0.071753, 0.149705]$
On remarque que le sommet 6 à la plus grande proportion mais aussi que le sommet 2, 3 et 8 sont relativement proche

Graphe après transformation



Quand on fait tourner backspace sur le graphe on obtient cela :
Pour $\alpha = 0.85$:
$x = [0.027459, 0.075200, 0.194179, 0.048690, 0.020492, 0.412774, 0.027027, 0.194179]$

On remarque bien que le sommet 6 à la plus grande proportion ce qui semble logique étant donné la redistribution des nouveaux arcs vers 6.

Les résultats sont tout à fait logique car avec le mécanisme de backspace les sommets menant vers des pages sans liens de sortie vont avoir une probabilité significativement plus grande d'être choisie à l'étape d'après alors qu'en tant normale ils auront une probabilité uniforme d'être choisis.

4 D'autres exemples plus grands

Nous allons maintenant tester cette approche sur des graphes de taille plus importante afin de déterminer si la convergence nécessite davantage d'itérations.

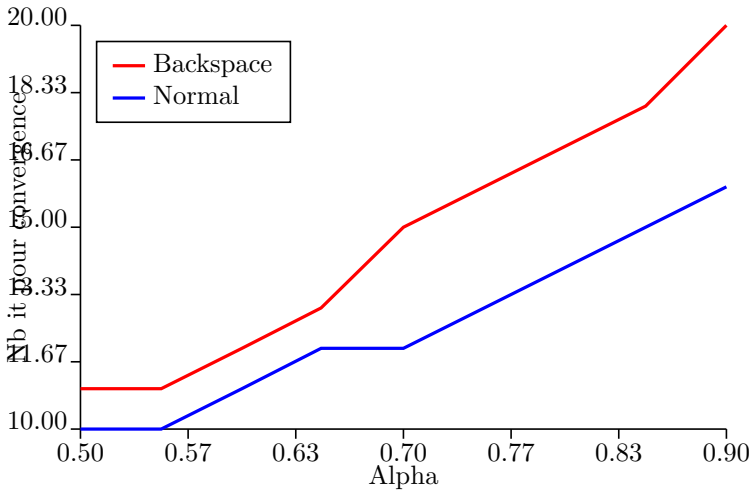


Figure 1: Graphe illustrant la convergence d'une matrice à 10 001 sommets avec 1% de sommets de degré sortant nul.

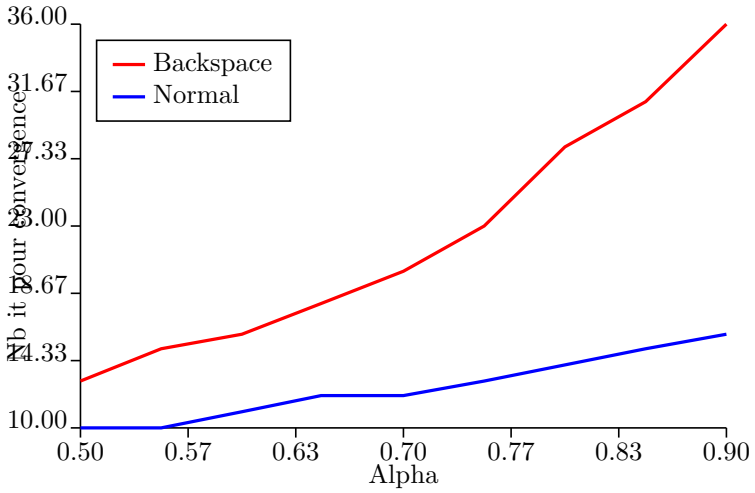


Figure 2: Graphe illustrant la convergence d'une matrice à 10 001 sommets avec 10% de sommets de degré sortant nul.

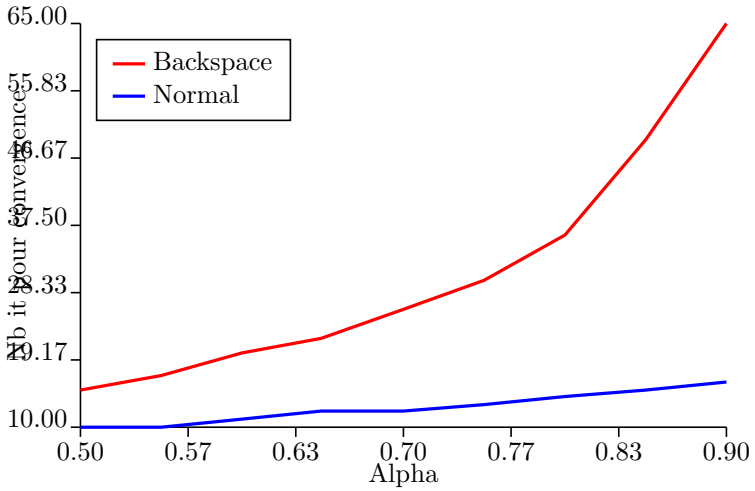


Figure 3: Graphe illustrant la convergence d'une matrice à 10 001 sommets avec 25% de sommets de degré sortant nul.

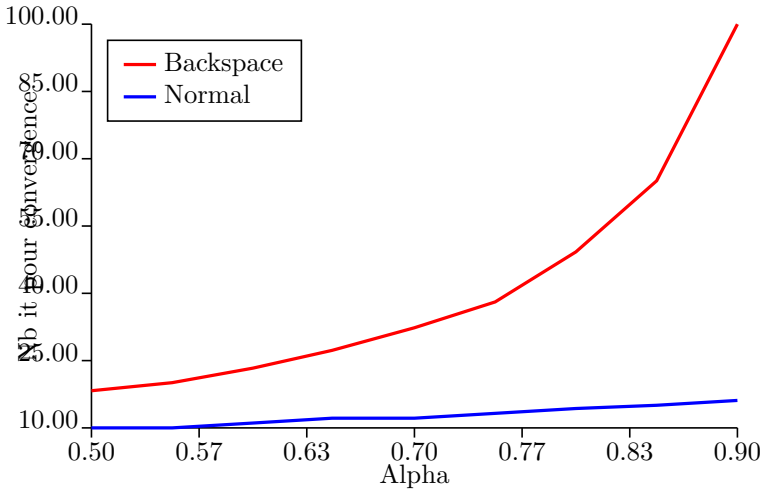


Figure 4: Graphe illustrant la convergence d'une matrice à 10 001 sommets avec 50% de sommets de degré sortant nul.

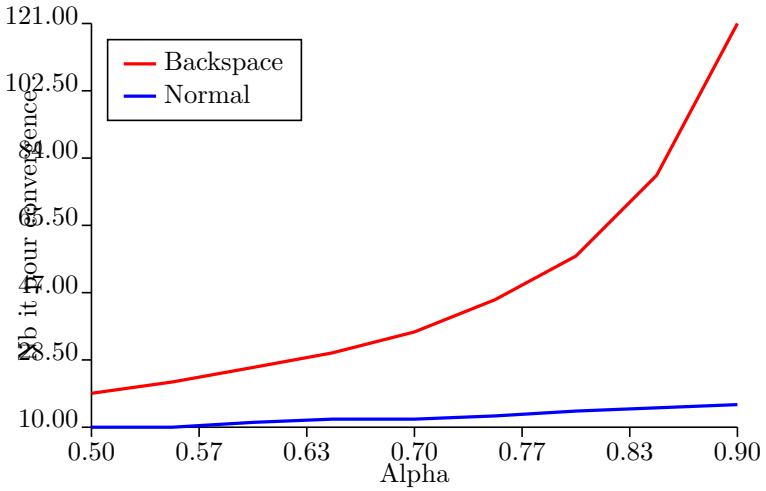


Figure 5: Graphe illustrant la convergence d'une matrice à 10 001 sommets avec 75% de sommets de degré sortant nul.

Sur ces graphiques, on observe que la convergence est de plus en plus lent en fonction du nombre de sommets avec un degré sortant nul dans le graphe. En effet, plus le pourcentage de sommets avec un degré sortant nul est grand et plus le Backspace met du temps a converger (jusqu'à être plus de 10 fois plus lent que le pagerank classique, debute à 20 itérations pour alpha = 0.9 à 121 itérations) alors que le pagerank garde ça vitesse constante (environ 15 iterations pour alpha = 0.9).

Cette différence s'explique par la gestion des sommets de degré sortant nul, qui sont alors divisés en plusieurs sommets reliés par des arcs de probabilité égale à 1. Or, répartir une probabilité de 1 sur le voisinage prend beaucoup plus de temps avant d'atteindre l'état stationnaire. C'est pourquoi il est préférable d'initialiser la distribution stationnaire en répartissant uniformément la probabilité sur tous les états, plutôt que de donner immédiatement une probabilité de 1 à un seul sommet et 0 aux autres.

5 Comparaison des distributions stationnaires pour $\alpha = 0{,}85$

Nous présentons ci-dessous les vecteurs stationnaires obtenus par les méthodes Backspace et PageRank classique, en fonction du pourcentage de sommets avec un degré sortant nul dans le graphe.

20 % de sommets avec un degré sortant nul

- **Backspace** : $[0{,}1196,\ 0{,}1611,\ 0{,}0687,\ 0{,}1344,\ 0{,}0889,\ 0{,}1230,\ 0{,}1962,\ 0{,}1081]$
- **PageRank classique** : $[0{,}1396,\ 0{,}1581,\ 0{,}0764,\ 0{,}1360,\ 0{,}1092,\ 0{,}1012,\ 0{,}1849,\ 0{,}0890]$

Les deux distributions sont très proches : l'impact des sommets avec un degré sortant nul reste faible à 20 %.

50 % de sommets avec un degré sortant nul

- **Backspace** : $[0{,}1104,\ 0{,}1187,\ 0{,}1272,\ 0{,}2027,\ 0{,}0730,\ 0{,}2346,\ 0{,}0820,\ 0{,}0513]$
- **PageRank classique** : $[0{,}1396,\ 0{,}1581,\ 0{,}0764,\ 0{,}1360,\ 0{,}1092,\ 0{,}1012,\ 0{,}1849,\ 0{,}0890]$

La similarité persiste à 50 % de sommets avec un degré sortant nul : la méthode Backspace reste proche du PageRank standard.

80 % de sommets avec un degré sortant nul

- **Backspace** : $[0{,}3921,\ 0{,}0189,\ 0{,}0684,\ 0{,}3552,\ 0{,}1092,\ 0{,}0188,\ 0{,}0188,\ 0{,}0188]$
- **PageRank classique** : $[0{,}1396,\ 0{,}1581,\ 0{,}0764,\ 0{,}1360,\ 0{,}1092,\ 0{,}1012,\ 0{,}1849,\ 0{,}0890]$

À 80 %, on note une divergence importante : Backspace concentre massivement la probabilité sur un seul sommet ($\approx 0,39$), tandis que PageRank classique reste équilibré.

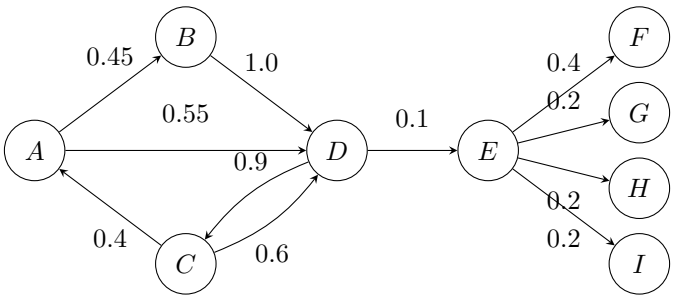
Interprétation

Cette différence s'explique par le traitement des sommets avec un degré sortant nul, qui sont décomposés en nœuds reliés par des arcs de probabilité unitaire : redistribuer une probabilité de 1 sur leurs voisins allonge le temps de convergence et accentue la "richesse" d'un unique sommet. C'est pourquoi il est préférable d'initialiser la distribution stationnaire de façon uniforme sur tous les états.

6 Évaluation de la notation des pages

Après avoir examiné les distributions stationnaires précédentes, voyons maintenant si une page pointant vers un grand nombre de sommets avec degré nul (pages sans lien sortant) se classe différemment avec la méthode Backspace par rapport au PageRank classique.

Considérons le graphe suivant :



Pour $\alpha = 0{,}85$, les scores PageRank sont :

- $D = 0{,}2919$
- $E = 0{,}0592$

Ce résultat est attendu : D reçoit de nombreux liens à haute probabilité, tandis que E n'est pointé que par D avec 0,10.

Avec Backspace, on obtient :

- $E = 0{,}3075$
- $D = 0{,}1413$

Ici, E semble beaucoup plus pertinent, alors qu'il ne reçoit qu'un unique lien de D (0,10) et pointe lui-même vers plusieurs sommets de degré sortant nul.

Interprétation.

Le PageRank vise à classer les pages selon le nombre et la qualité des liens entrants ; or, Backspace transforme chaque sommets avec degré sortant nul en un nœud qui renvoie vers son prédécesseur avec probabilité 1, ce qui avantage artificiellement les pages qui citent de nombreux sommets avec degré sortant nul. Ce comportement s'oppose à l'esprit du PageRank original, où la notoriété provient essentiellement des citations de pages elles-mêmes influentes.

7 Conclusion

Ce travail montre que la variante "Backspace" du PageRank, qui réinjecte la probabilité des sommets de dégré sortant nul vers leur prédécesseur immédiat, modifie profondément à la fois la convergence numérique et le classement final des pages. D'une part, la déconcentration des probabilités unitaires vers les nœuds parents rallonge significativement le nombre d'itérations nécessaires, jusqu'à plus de dix fois comparé au PageRank classique pour de forts taux de sommets de degré nul. D'autre part, la redistribution systématique de la probabilité des sommets de degré nul entraîne un biais artificiel : les pages qui naviguent vers de nombreux sommets de degré nul "s'enrichissent" injustement, au détriment de celles réellement plébiscitées par un grand nombre de liens entrants.

En pratique, si l'objectif est d'imiter fidèlement le comportement d'un internaute qui utilise le bouton "précédent", la méthode Backspace reste pertinente. En revanche, pour conserver l'esprit originel du PageRank — à savoir valoriser les pages les plus citées par leurs pairs — il est préférable de maintenir une distribution uniforme sur les états ou d'adopter des traitements alternatifs des sommets de degré nul (par exemple, une téléportation partagée). Cette étude invite donc à réfléchir aux compromis entre réalisme du modèle de navigation et robustesse du classement obtenu.