SUJET D'ANALYSE

Etude du fonctionnement de l'algorithme de classement des pages Web utilisé par le moteur de recherche Google

Présenté par :
Kouamé Gérard Kra
Seyed Hossein Seyedi Nahrmiani
Koffi Roland Wotobe

Encadrant :
M Rubenthaler


PageRank

Table des matières

INTRODUCTION

Depuis sa conception en 1998, Google a connu plusieurs améliorations et cela fait de lui aujourd'hui le moteur de recherche le plus utilisé au plan mondial. Cependant, la plupart des réformes de Google demeurent jusqu'à présent des secrets. Par contre, selon l'un des articles publiés par les fondateurs [1], le pilier de son succès est une judicieuse modélisation mathématique que nous retraçons ici.

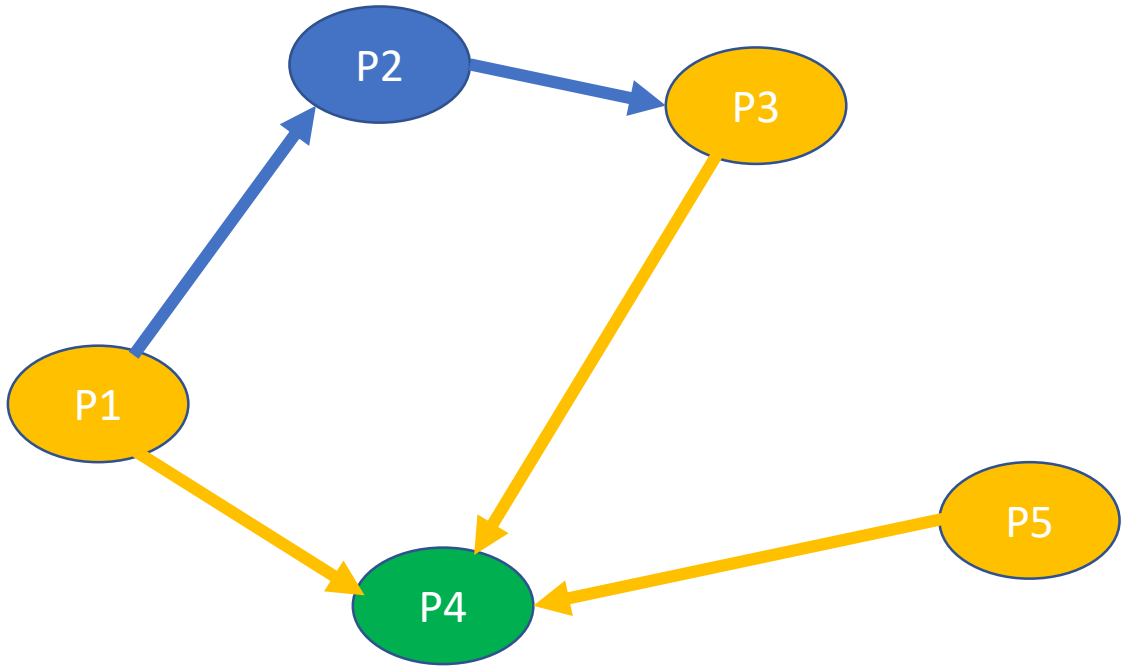
1. ORIGINE DE PAGERANK

Lorsqu'un utilisateur lance une recherche sur Google, comment procède Google pour déterminer l'ordre selon lequel présenter les résultats à l'utilisateur?

L’un des plus important algorithme utilisé par Google pour ce classement est PageRank. C’est un algorithme qui estime l’importance des pages Web.

En effet, une première approche consiste à dire qu’une page Web est importante si elle est citée par plusieurs autres pages Web.

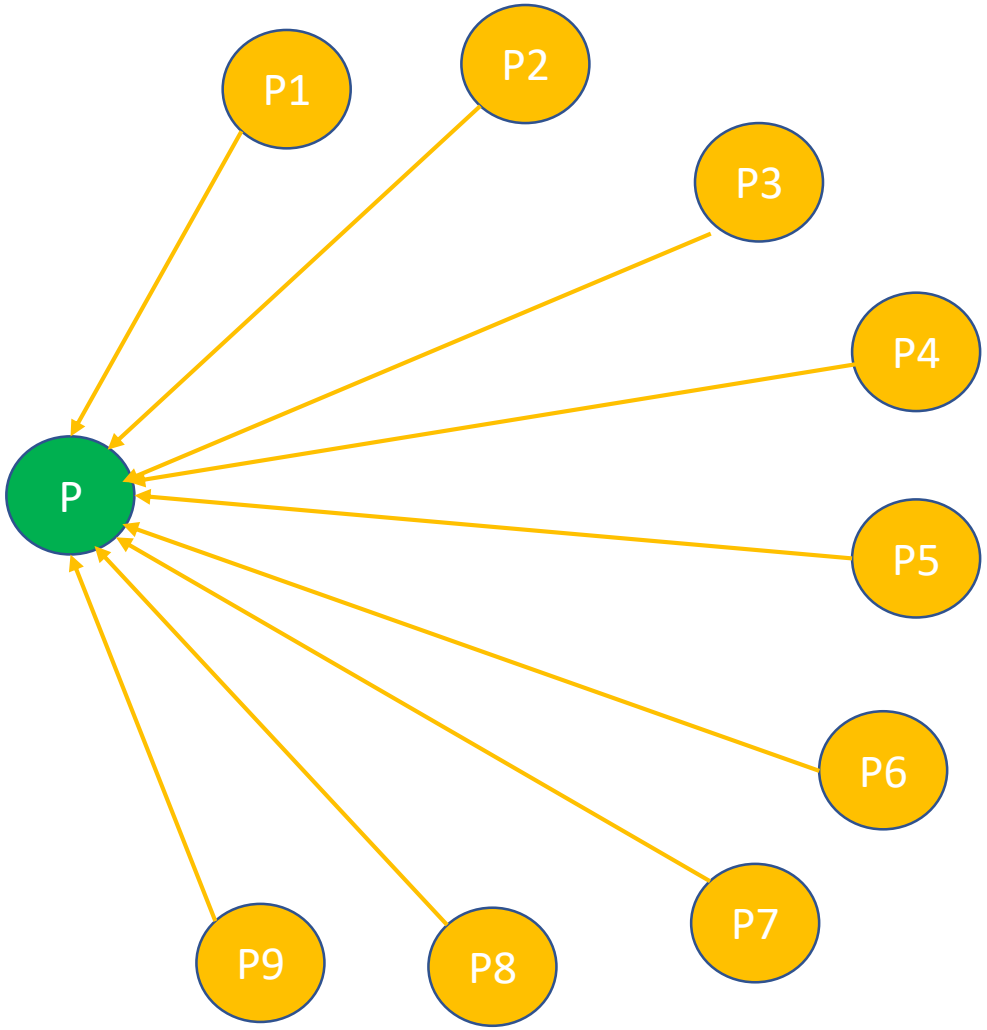
Sur la base de ce principe, nous pouvons dire dans l'exemple ci-dessous que la page P4 est la plus importante.



Il y a cependant un problème!

Cette seule approche ne suffit pas. En effet, partant de ce principe, un développeur Web peut facilement faire remarquer l'importance de sa page Web, en créant plusieurs autres pages qui citent la page en question.

Cela ressemblerait à l'exemple ci-dessous.


P1
P2
P3
P4
P
P5
P6
P7
P9
P8

C’est pourquoi une deuxième approche sera prise en compte: on dira qu’une page Web est importante si elle est citée par une autre page importante.

2. Description du fonctionnement de l’algorithme

Comment pouvons nous estimer l’importance d’une page sachant l’importance de la page qui la cite?

Modèle du surfer aléatoire

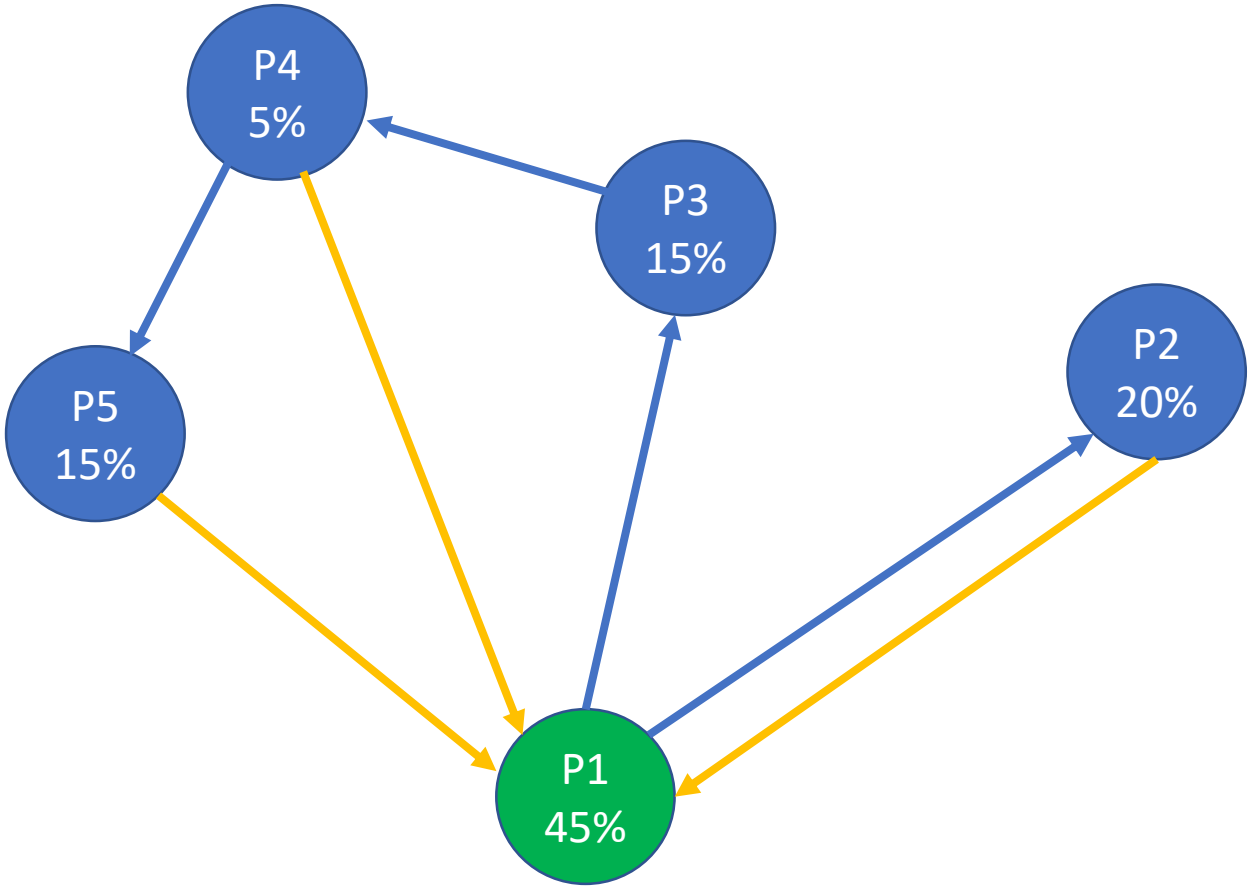
Considérons un surfer aléatoire qui se promène sur différentes pages Web en prenant de façon aléatoire un lien quelconque.

Le principe du modèle consiste à attribuer un score à la page chaque fois qu’elle sera visitée par le surfer puis l’incrémenter à chaque passage.

Après plusieurs itérations, la page ayant le plus gros score sera considérée comme la plus importante.

Il est évident que les pages pour lesquelles il existe plusieurs liens menant vers elles auront tendance à remporter plus de score que celles pour lesquelles il y a peu de lien menant vers elles.

Dans l’exemple ci-dessous, P1 est la page la plus importante.



Il y a par contre un souci pour l'approche ci-dessus!

L'approche ci-dessus nous donne l'impression que le surfer aléatoire reste uniquement dans un ensemble de pages Web interconnecté, ce qui n'est pas exactement le cas dans le monde réel.

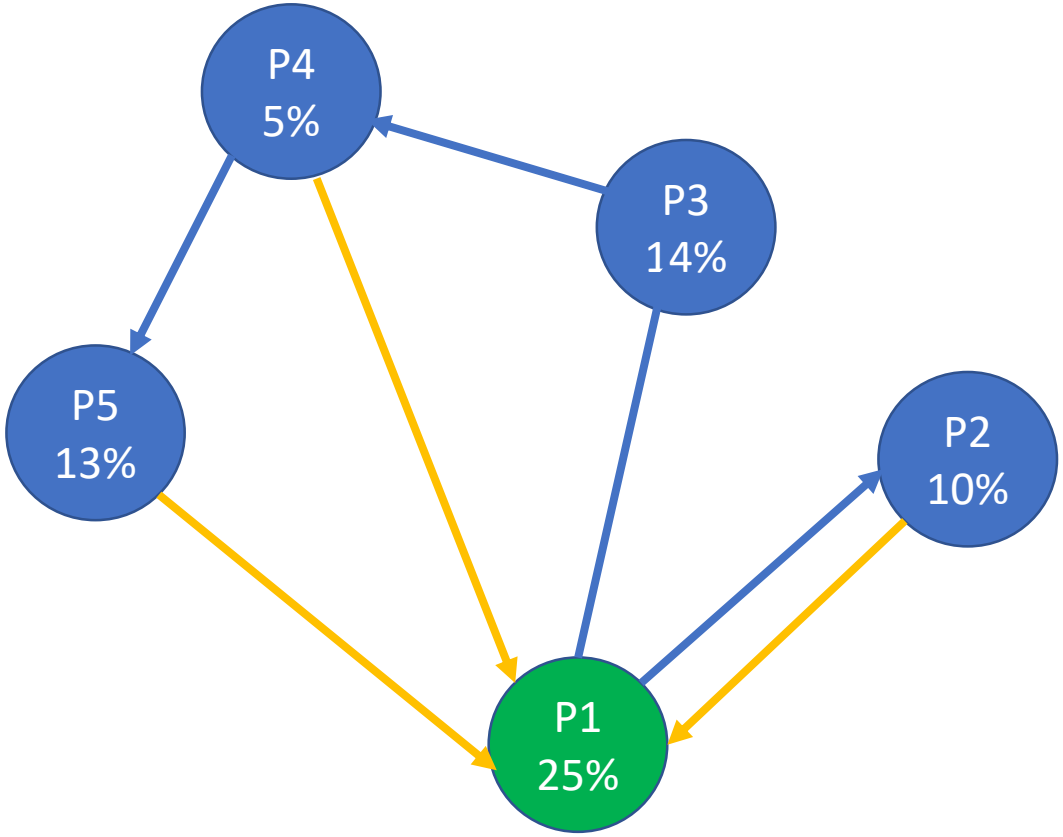
En effet, il peut arriver qu'un ensemble de pages Web ne soit liée à un ou plusieurs autres ensemble de pages Web. Comment pouvons-nous prendre en compte cet aspect?

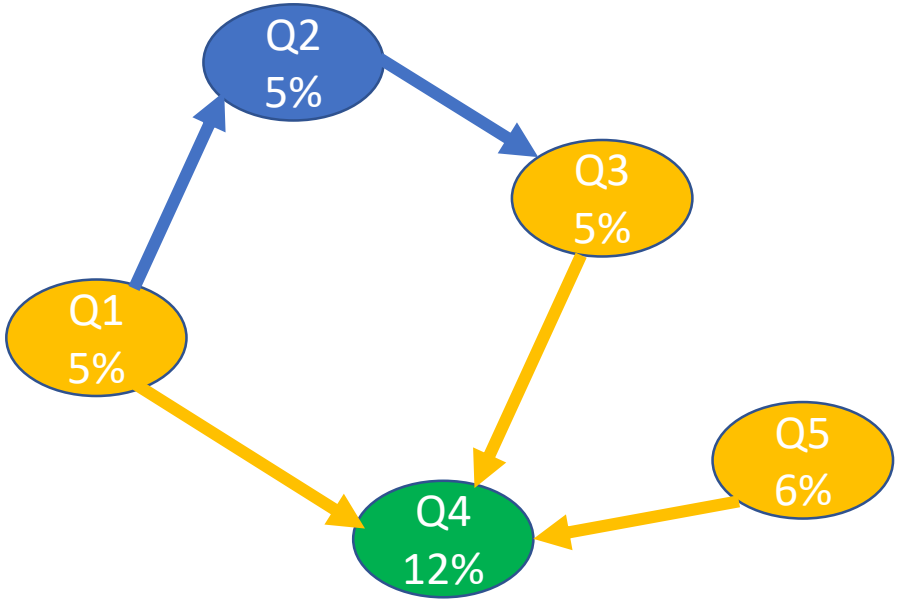
2.2. Damping Factor

La resolution du problème exposé ci-dessus nécessite l'introduction de la notion de Damping Factor (en anglais) ou Facteur d'Amortissement (en Français).

Ce facteur modélise le comportement des utilisateurs qui pourraient simplement naviguer de manière aléatoire sans suivre de liens.

Pour un Damping Factor d=0.15

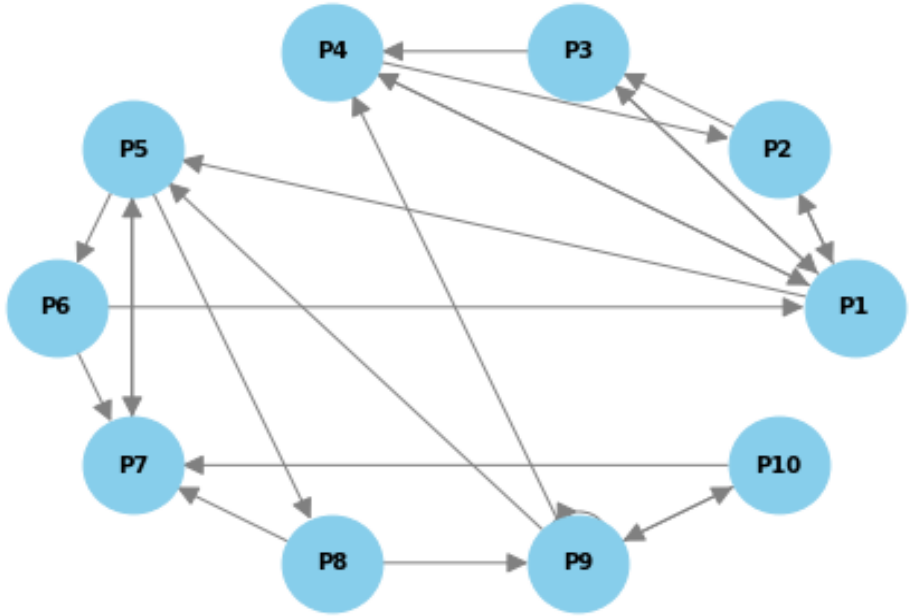




3. Mise en œuvre

Le modèle PageRank est donnée par la formule $p'_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{l_j} \times p_j$

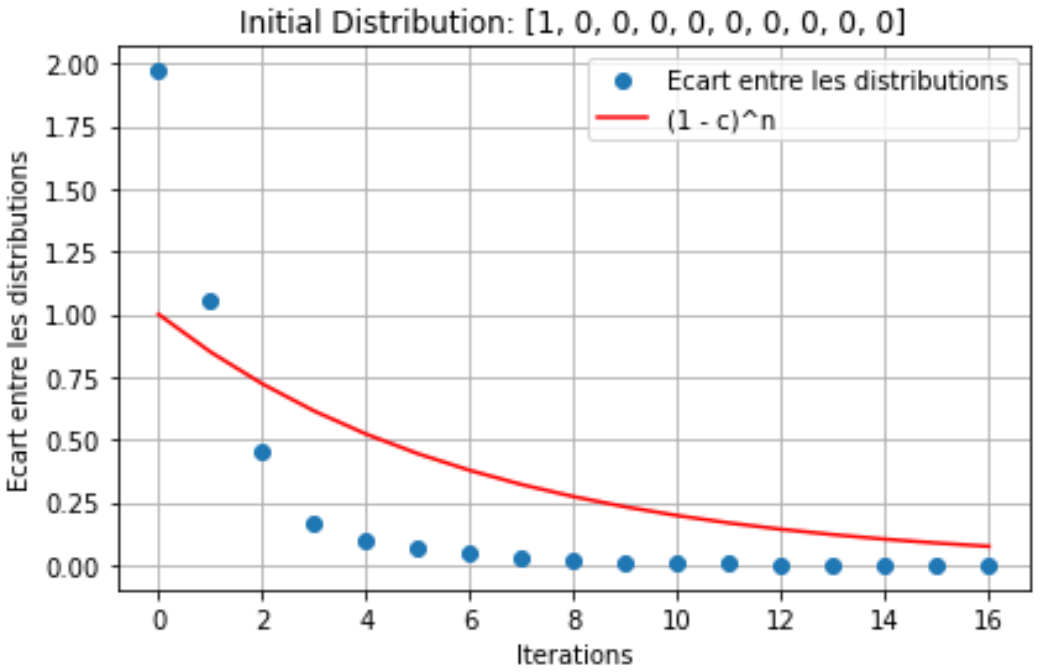
Où c représente le Damping factor et le poids associé à la quantité (1-c) représente la matrice de transition associée au graphe du Web

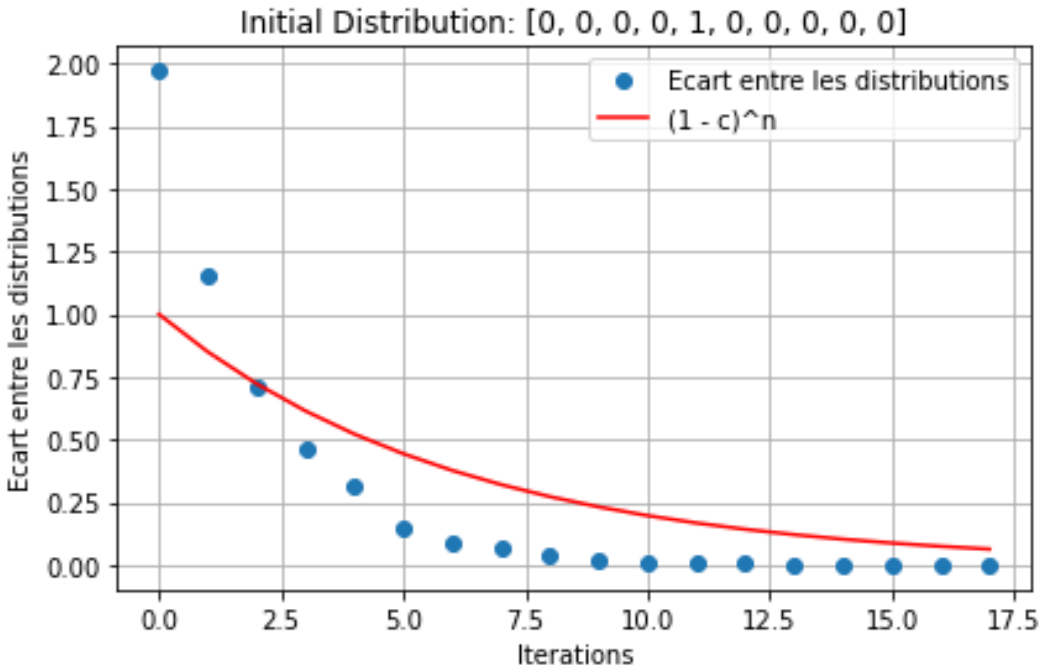


Exemple de graphe du Web avant application du modèle

Pour un Damping Factor c=0.15

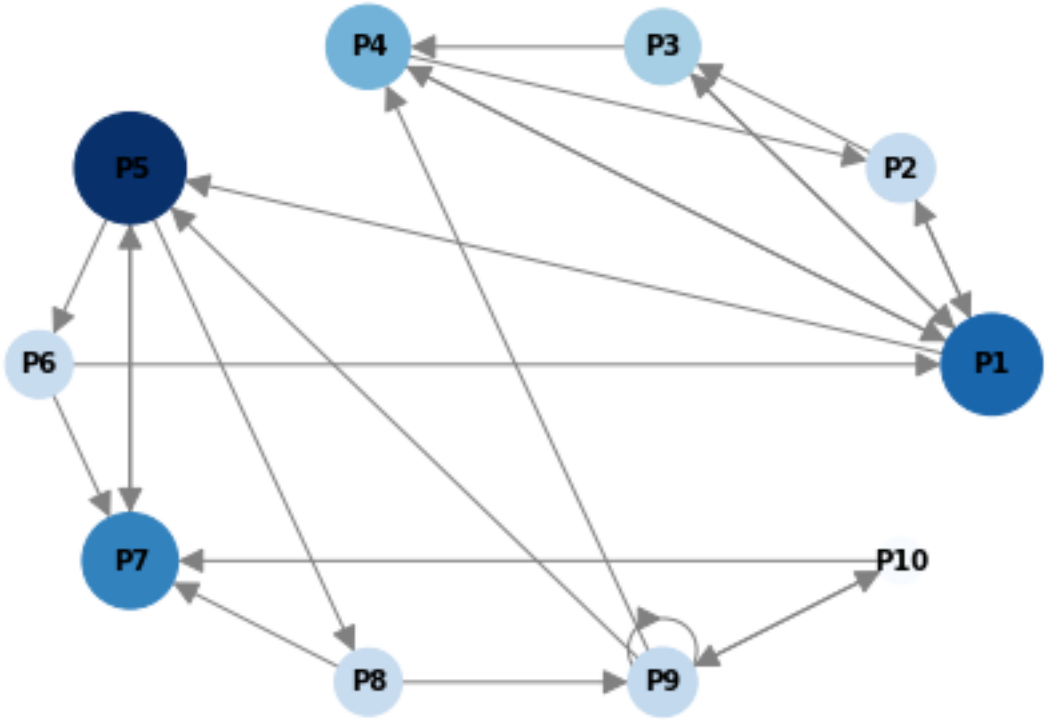
$$p'_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{l_j} \times p_j$$





Pour un Damping Factor c=0.15

$$p'_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{l_j} \times p_j$$



Le graphe du Web après application du modèle

CONCLUSION

PageRank, n’est pas le seul moyen par lequel calculer l’importance d’une page Web. Cependant, son approche a révolutionné le domaine de la recherche sur Internet en fournissant une méthode efficace et pertinente pour évaluer la pertinence des pages web et fournir à l’utilisateur les résultats de sa recherche par ordre d’importance.

Référence

[1] *S. Brin, L. Page The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford University 1998,*

http://images.math.cnrs.fr/Comment-Google-classe-les-pages.html?lang=fr

https://youtu.be/meonLcNCours de ETH Zurick : Chapter 11
https://disco.ethz.ch/courses/ti2/ 7LD4

Merci pour votre attention!