La matrice cachée de Google

[1]

Note Ce document ne vise en aucun cas à décrire exhaustivement le travail mis en oeuvre pour l'épreuve du TIPE, mais plutôt décrire le cheminement de pensée qui a amené à sa conception.

Problématique

Comment les concepteurs de Google font-ils pour classer des milliers de pages se rapportant à un mot-clé donné, de façon telle que les pages les plus représentatives occupent les premières positions du classement ?

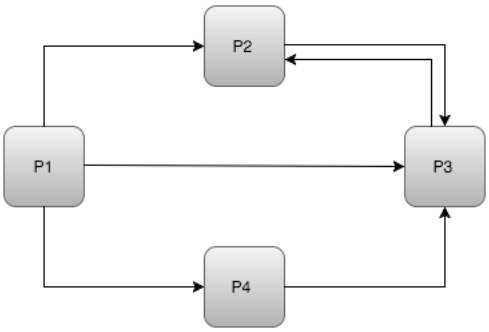
Table des matières

1. Épreuve de TIPE 2012-2013, thème Invariance et similitude : La matrice de Google, par RIHANI Mohammed

1 Principe de l'algorithme PageRank

1.1 Modèle inital (matrice élémentaire)

Nous allons présenter, au moyen d'un exemple de taille réduite, le problème qui nous intéresse : classer des pages pour déterminer les plus significatives par rapport à un sujet donné.



Pour classer ces pages, on leur attribue un score, aussi appelé PageRank. Plus le score est élevé, plus la page est importante. L'idée à la base du modèle de S.Brin et L.Page tient en deux règles :

- `R1` On accorde un PageRank plus élevé, aux pages référencées par des pages qui font elle-même autorité dans le domaine, c'est à dire qui ont un PageRank élevé.
- `R2` On accorde d'autant moins de crédit à une référence, si elle provient d'une page qui dispose de nombreux liens.

Les deux règles sont somme toute assez naturelles. Pour la première, au plus une page est référencée par d'autres pages, au plus elle doit faire autorité dans le domaine en question.
La deuxième règle sert de contrepoids : on ne peut qu'accorder moins de poids, aux sites qui gaspillent leurs recommandations.

Si je vous demande par exemple de citer les plus grands scientifiques de tous les temps. Les premiers du classement seront logiquement les scientifiques les plus cités. Cependant, accordez vous la même importance à la liste fournie par X et qui contient un unique nom "Albert Eintein" et la liste fournie par Y qui contient aussi "Albert Einstein", mais également d'autres scientifiques.
Clairement, X considère "Albert Einstein" comme le plus grand scientifique de tous les temps. Par contre pour Y, il ne s'agit que l'un des plus grands. Cet exemple atteste parfaitement du rôle de la deuxième règle.
La première règle quant à elle voudrait qu'on accorde plus de poids pour la liste fournie par un prix Nobel ou par un scientifique de renom que pour celle d'un citoyen *lambda* sans qualifications scientifiques parliculières.

Reprenons notre exemple et appelons $S1$, $S2$, $S3$ et $S4$ les PageRank des pages 1 à 4. Une prise en compte de la règle `R1` nous donne le système :

$$\begin{cases} S1 &= 0 \\ S2 &= S1 + S3 \\ S3 &= S1 + S2 + S4 \\ S4 &= S1 \end{cases}$$

De plus si on envisage `R2`, et on divise le poids attribué aux liens de chaque page par le nombre de liens de celle-ci, le système devient :

$$\begin{cases} S1 &= 0 \\ S2 &= \frac{S1}{3} + S3 \\ S3 &= \frac{S1}{3} + S2 + S4 \\ S4 &= \frac{S1}{3} \end{cases}$$

La matrice élémentaire de Google est la matrice du système. Dans notre exemple :

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 1 & 0 \\ \frac{1}{3} & 1 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix}$$

Ainsi on a le système :

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 1 & 0 \\ \frac{1}{3} & 1 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} S1 \\ S2 \\ S3 \\ S4 \end{pmatrix} = \begin{pmatrix} S1 \\ S2 \\ S3 \\ S4 \end{pmatrix}$$

En fait, à ce stade, deux problèmes majeurs se posent. Rien ne garantit que le système admette au moins une solution non nulle. Ensuite, si le système possède effectivement une solution (non nulle), il serait intéressant d'en garantir l'unicité. En effet, si plusieurs solutions sont disponibles, comment donner du sens à la solution qui sera calculée si d'autres solutions, tout aussi valables, mais différentes, peuvent êtres trouvées. On aurait alors plusieurs classements incomparables...

1.2 Recherche d'une solution et perturbation du modèle

— Remplacer les colonnes nulles de M par des colonnes dont tous les éléments sont égaux à $\frac{1}{n}$ (n étant la dimension de la matrice).

— $\boxed{G = \alpha M + \frac{(1-\alpha)}{n}.J}$ (où J est la matrice dont tous les éléments sont égaux à 1 et $\alpha = 0.85$).

Remarque Le choix de la valeur $\alpha = 0.85$ n'est pas arbitraire et constitue en fait un bon compromis : plus la valeur de α est proche de 1, plus on est proche du modèle initial (si $\alpha = 1$, alors $G = M$), mais d'un autre côté pour des raisons de rapidité et de stabilité des calculs réalisés, il vaut mieux ne pas choisir une valeur trop proche de 1.

Les constructions réalisées permettent d'assurer à G d'être primitive et stochastique et donc nous donne droit à l'application du théorème de Perron-Frobenius.

Théorème de Perron-Frobenius Soit $A = (a_{ij})$ de $M_n(R)$ une matrice positive et $r(A)$ son rayon spectral.

- `i) Perron` : si A est primitive alors $r(A) > 0$ et $r(A)$ est une valeur propre dominante et simple de A et
  $\exists!$ vecteur $x > 0$ tel que $Ax = r(A).x$ et $|x| = \sum x_i = 1$
- `ii) Frobenius` : si A est irréductible on a les mêmes résultats que Perron excepté que $r(A)$ ne sera pas nécessairement dominante.

De plus on fait appelle aux propriétés :

- `a)` Le rayon spectral d'une matrice stochastique M est égal à 1, $r(M) = 1$
- `b)` Si M est stochastique telle que la valeur propre $r(M) = 1$ est simple et dominante alors la suite (M^n) converge, et pour tout vecteur de probabilité X (c'est à dire vecteur stochastique), la suite $(M^n X)$ converge vers l'unique vecteur de pobabilité invariant de M.

Ces propriétés mathématiques engendrent l'existence et l'unicité d'un classement vérifiant : $Gs = s$.
En outre, les propriétés `a)` et `b)` entrainent également que la suite (G^n) converge vers une matrice limite dont les colonnes sont toutes égales à s (le vecteur des PageRank recherché).

Ainsi, on retrouve l'aspect de l'invariance qui se manifeste dans la matrice de Google G. Ces puissances deviennent invariantes à partir d'un certain rang.

Revenons à notre exemple, après calcul :

$$G = \begin{pmatrix} \frac{3}{80} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{71}{80} & \frac{3}{80} \\ \frac{77}{240} & \frac{71}{80} & \frac{3}{80} & \frac{71}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \end{pmatrix}$$

$$G^5 = \begin{pmatrix} 0.0375 & 0.0375 & 0.0375 & 0.0375 \\ 0.52308 & 0.22727 & 0.67098 & 0.022727 \\ 0.39129 & 0.68709 & 0.24339 & 0.68709 \\ 0.04812 & 0.04812 & 0.04812 & 0.04812 \end{pmatrix}$$

$$G^{20} = \begin{pmatrix} 0.0375 & 0.0375 & 0.0375 & 0.0375 \\ 0.43940 & 0.46534 & 0.42649 & 0.46524 \\ 0.47496 & 0.44912 & 0.48788 & 0.44912 \\ 0.04812 & 0.04812 & 0.04812 & 0.04812 \end{pmatrix}$$

$$G^{80} = \begin{pmatrix} 0.0375 & 0.0375 & 0.0375 & 0.0375 \\ 0.44613 & 0.44613 & 0.44613 & 0.44613 \\ 0.46824 & 0.46824 & 0.46824 & 0.46824 \\ 0.04812 & 0.04812 & 0.04812 & 0.04812 \end{pmatrix}$$

$$G^{200} = \begin{pmatrix} 0.0375 & 0.0375 & 0.0375 & 0.0375 \\ 0.44613 & 0.44613 & 0.44613 & 0.44613 \\ 0.46824 & 0.46824 & 0.46824 & 0.46824 \\ 0.04812 & 0.04812 & 0.04812 & 0.04812 \end{pmatrix}$$

1.3 Interprétation probabiliste

Le caractère probabiliste de la matrice G se manifeste dans le fait que l'élément se trouvant en j-ième ligne et i-ième colonne encode la probabilité, lorsque le surfeur se trouve sur la page p_i, de visiter la fois suivante la page p_j, $P(p_i \to p_j)$.

$$G = \begin{pmatrix} P(p_1 \to p_1) & P(p_2 \to p_1) & P(p_3 \to p_1) & P(p_4 \to p_1) \\ P(p_1 \to p_2) & P(p_2 \to p_2) & P(p_3 \to p_2) & P(p_4 \to p_2) \\ P(p_1 \to p_3) & P(p_2 \to p_3) & P(p_3 \to p_3) & P(p_4 \to p_3) \\ P(p_1 \to p_4) & P(p_2 \to p_4) & P(p_3 \to p_4) & P(p_4 \to p_4) \end{pmatrix}$$

Se trouvant sur une page donnée, le surfeur a deux options :

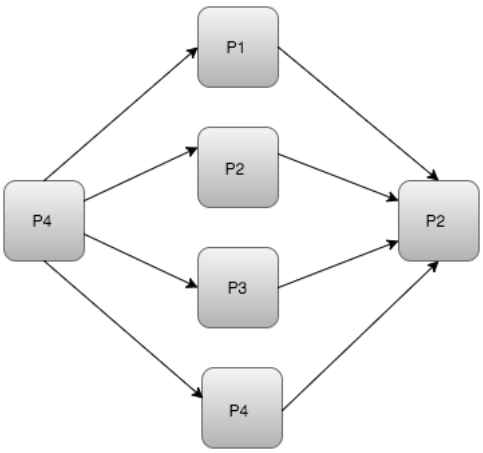
- `a)` Avec une probabilité de 0.85, il choisit aléatoirement un des liens présents sur la page actuellement visitée.
- `b)` Avec une probabilité de 0.15, il est redirigé aléatoirement vers des pages ayant une même probabilité du web (chacunes des pages ayant une même probabilité $\frac{1}{n}$ d'être choisie lors de cette redirection, si n représente le nombre de pages web de l'internet).

Mais que représente G^2 en ces termes probabilistes ?
Pour répondre à cette question, il faut repenser à la définition même du produit matriciel. Si l'on désir obtenir l'élément se trouvant à la deuxième ligne et à la quatrième colonne de G^2, on obtient :

$G_{21}G_{14} + G_{22}G_{24} + G_{23}G_{34} + G_{24}G_{44} = P(p_1 \to p_2)P(p_4 \to p_1) +$
$P(p_2 \to p_2)P(p_4 \to p_2) + P(p_3 \to p_2)P(p_4 \to p_2) + P(p_4 \to p_2)P(p_4 \to p_4)$

Ce nombre représente exactement la probabilité qu'a le surfeur de se retrouver en deux unités de temps (i.e avec un chemin de longueur 2) à la page 2 s'il est parti de la page 4.

Par exemple, le produit $P(p_4 \to p_1)P(p_1 \to p_2)$ représente la probabilité d'aller :



Si on poursuit ce raisonnement, on obtient (par une simple récurrence sur n) que G^n encode les probabilités de transitions entre pages après n coups.

2 Application PageRank (Championnats sportifs)

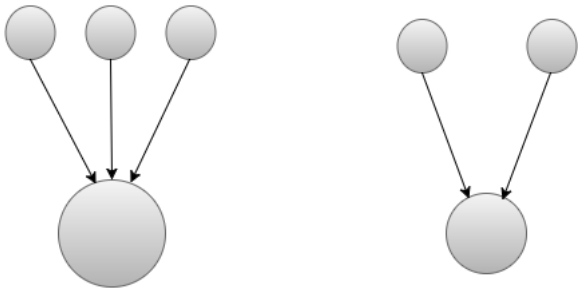
La technique développée pour classer les pages peut s'appliquer à d'autres situations comme un championnat sportif. On peut prendre le basket ou le tennis, mais pas le football, car il peut y'avoir des matchs nuls en ce dernier.
Ainsi, un match est toujours gagné ou perdu. Si une équipe A (ou un joueur A pour le tennis) bat une équipe B, on trace un arc allant de B vers A. Les deux règles `R1` et `R2` s'adaptent parfaitement aux championnats sportifs tels que le basket ou le tennis.

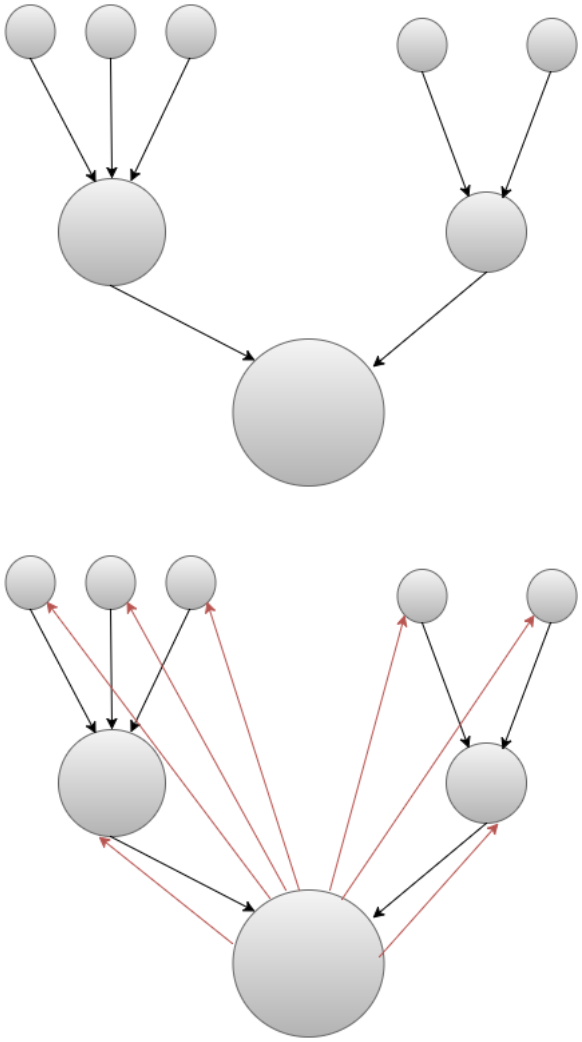
— `R1` On accorde d'autant plus d'importance aux matchs gagnés contre une équipe ou un joueur réputé fort, c'est à dire qui possède un score élevé.

— `R2` On accorde d'autant moins d'importance aux matchs gagnés contre une équipe ou un joueur qui perd toutes ses rencontres ou presque.

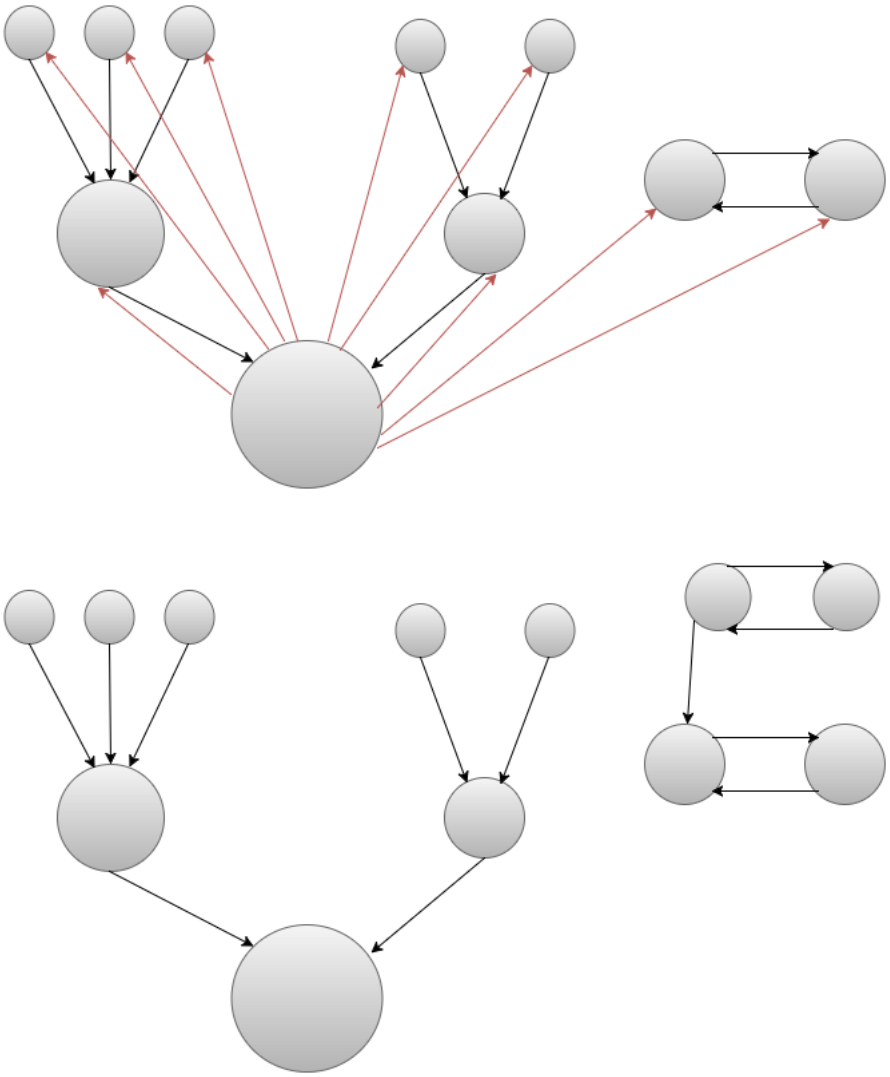
3 Analyse PageRank

3.1 Procédés pour augmenter le PageRank





Link farms (Fermes de liens)



3.2 Conclusions et perspectives

Références

— "Les matrices : théorie et pratique" de D.Serre, Dunod 2001.
— "Google's PageRank and beyond : the science of search engine rankings" de A.N. Langville et C. D. Meyer, Princeton University Press 2006."
— L'algorithme PageRank de Google : une promenade sur la toile de Michael Eisermann
— http ://www.discmath.ulg.ac.be/ .
— http ://www.pagerank.com/ .