



SEO Campus 2009 : Pagerank et optimisation

Sylvain Peyronnet

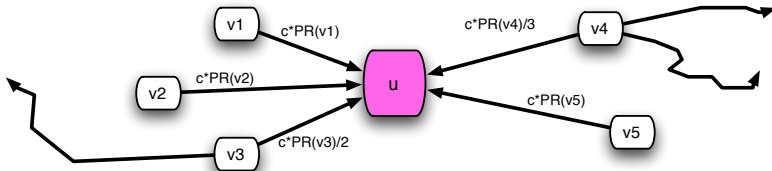
<http://sylvain.berbiqui.org>

<http://www.kriblogs.com/syp>

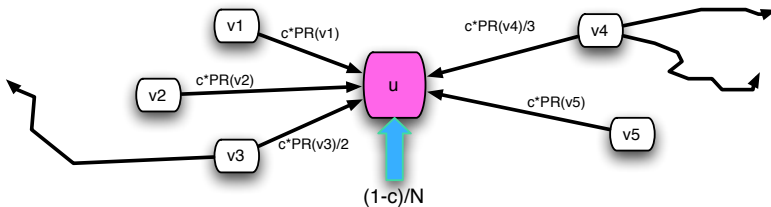
04/02/2009



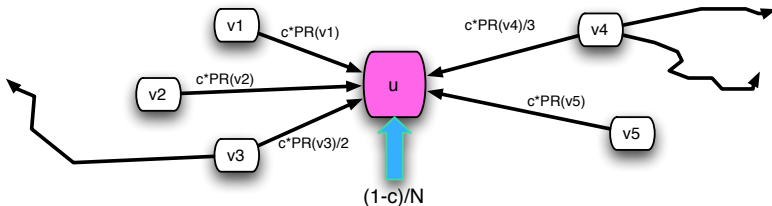
PageRank : la vision classique



PageRank : la vision classique



PageRank : la vision classique



Initialisation : $\forall u \ PR(u) = 1/N$

Calcul itératif :

$$PR(u) = \frac{(1-c)}{N} + c \cdot \sum_{v \rightarrow u} \frac{PR(v)}{\#liens(v)}$$

Concept classique : Mark (1988), Bray (1996), Marchiori (1997), Brin et Page (1998), Kleinberg (1999) ...



PageRank : la vision classique

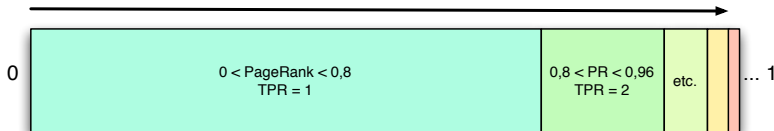
A retenir :

$$\forall u \quad 0 < PageRank(u) < 1$$

$$\sum PageRank(u) = 1$$

Le PageRank est intéressant car c'est une notion simple et facile à calculer

Relation au Toolbar PageRank (TPR) :





PageRank : le surfeur aléatoire

Considérons le comportement suivant d'un internaute :

1. Tirer une page web au hasard
2. Tirer un nombre p entre 0 et 1
3. Si $p > c$ alors choisir une nouvelle page au hasard
4. si $p < c$ choisir au hasard un lien de la page web et aller à la page liée par ce lien (si pas de lien : goto 1)

La probabilité que cet internaute se trouve en une page donnée à un moment donné est égale au PageRank de cette page.

En conséquence : fort PageRank = forte probabilité d'être visité

Pourquoi maximiser le PR d'une page ?

Un moteur de recherche = 2 tâches

Pertinence
(Salton, tf*idf etc.)

Classement global
(PageRank, filtres)



Maximiser le PR d'une page

Pourquoi maximiser le PR d'une page ?

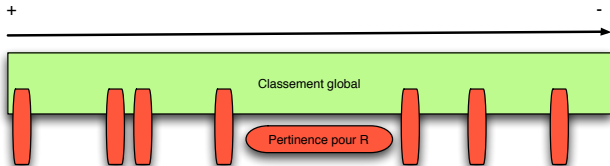
Un moteur de recherche = 2 tâches

Pertinence
(Salton, tf*idf etc.)

Classement global
(PageRank, filtres)

Classement sur une requête R :

$$\text{Classement}(R) = \text{Pertinence}(R) \cap \text{Classement global}$$





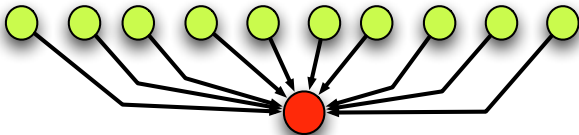
Maximiser le PR d'une page

Notion de ferme de liens (spam farm) :

Zoltán Gyöngyi, Hector Garcia-Molina. Link Spam Alliances. 31st International Conference on Very Large Data Bases (VLDB), 2005.

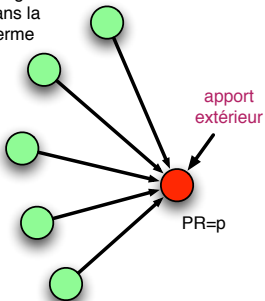
Donne la structure **optimale** pour maximiser le PageRank d'une page cible.

Structure classique (non optimale)



Maximiser le PR d'une page

k pages
dans la
ferme

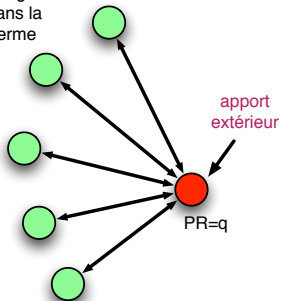


Structure simple

$$PR(\text{●})=p$$

PageRank
multiplié par
3,6
($c=0,85$)

k pages
dans la
ferme

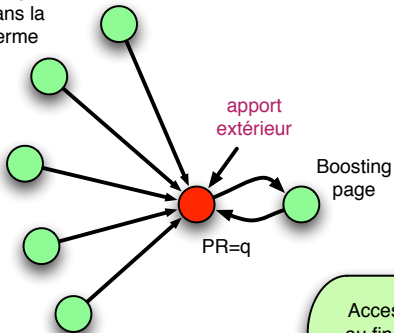


Structure optimale

$$PR(\text{●})=q=3,6 * p$$

Maximiser le PR d'une page

k pages
dans la
ferme



PageRank
multiplié par
3,6
($c=0,85$)

Accessibilité difficile !
au final PR plus faible
(indexation)

Structure optimale bis

$$PR(\bullet) = q = 3,6 * p$$

Optimalité :

1. Toutes les pages de la ferme pointent sur la cible (et pas ailleurs)
2. Le PR détourné pointe la cible (et pas ailleurs)
3. il y a des liens de la cible vers une ou plusieurs pages de la ferme

intuition : “guider” le surfeur aléatoire et le garder dans **notre** circuit.

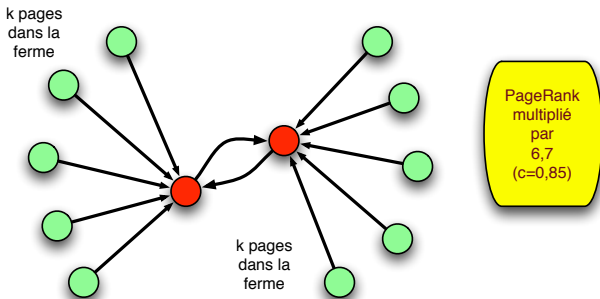
Stabilité :

Structure relativement stable à la perturbation



Maximiser le PR d'une page

Jouer en couple, ou plus ...



A plus : la stratégie de liens entre les cibles est importante et certaines fermes peuvent y perdre.

Est-ce detectable ?

Detecter la présence d'une ferme de liens au sein du web est un problème **NP-complet**.

Cela signifie qu'il **n'existe pas d'algorithme efficace pour detecter exactement la structure de ferme de liens**.

Détecter toutes les fermes de 50 pages : 2 mois sur Roadrunner
Détecter toutes les fermes de 100 pages : 100 milliards de milliards d'années (Roadrunner)

Des tentatives : Saito et al (2007), Bechetti et al (2008) ainsi que tous les algorithmes à base de Trust/Spam Rank.

Qu'est ce que le PageRank d'un site ?

C'est une notion qui n'a pas de sens mathématique à l'origine

La croyance populaire : c'est le PageRank de la page "principale" d'un site

Pour le théoricien :

$$\textit{PageRank}(\textit{site}) = \sum \textit{PageRank}(\textit{pages du site})$$

Pourquoi c'est important pour le SEO ?

Il faut penser au pauvre **surfeur aléatoire**...

Le PageRank c'est la probabilité de passage du surfeur aléatoire, dont le comportement est celui du visiteur "crétin".

fort PR = forte proba de passage

Le PageRank d'un site est donc la probabilité de passage sur le site (toutes pages confondues)



Maximiser le PR d'un site

Problème très étudié

1. Monica Bianchini, Marco Gori, Franco Scarselli: Inside PageRank. ACM Trans. Internet Techn. 5(1): 92-128 (2005)
2. Konstantin Avrachenkov, Nelly Litvak, Decomposition of the Google PageRank and optimal linking strategy, 2004.
3. Konstantin Avrachenkov, Nelly Litvak, The effect of new links on Google PageRank, Stoch. Models, 22 (2) (2006)
4. Amy N. Langville, Carl D. Meyer, Deeper inside PageRank, Internet Math. 1 (3) (2004) 335–380.

Mais problème réglé

C. de Kerchove, L. Ninove and P. Van Dooren. Maximizing PageRank by some Outlinks. Journal of Linear Algebra and its Application, vol. 429, September 2008.

cas 1 : trivial

C'est le cas où il n'y a pas nécessairement de liens vers l'extérieur.

Dans ce cas on fait un graphe complet entre les pages du site et c'est gagné. On peut obtenir cela en mettant tous les liens vers l'extérieur en **nofollow**.

cas 2 : difficile

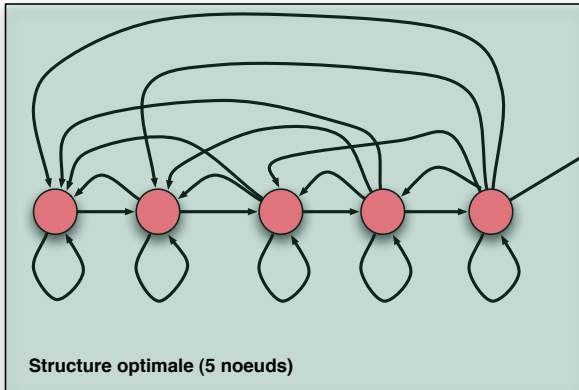
Chaque page doit avoir un accès à l'extérieur.

Cas très complexe, accès = un chemin qui mène à l'extérieur.
L'intuition va être de faire cycler au maximum le surfeur aléatoire.



Maximiser le PR d'un site

Structure optimale



Remarques :

1. La stabilité n'est pas claire (non stable selon Lempel et Moran, 2005)
2. Pour minimiser la fuite de PageRank, la page de sortie doit être celle avec le plus faible PageRank (= celle avec le moins de liens entrants ?).
3. Est-ce que cette structure est admissible pour un site ? (je laisse les webmasters répondre)



Questions ?

A quiz interface featuring a man's face on the left and a bar chart on the right. The bar chart has four bars of increasing height, labeled A, B, C, and D. Above the bars are percentages: 42%, 56%, 21%, and 0%. Below the bars is the formula $A+B+C+D$.

Qu'est-ce qui gravite autour de la Terre ?

- A: La Lune
- B: Le Soleil
- C: Mars
- D: Vénus