

Le PageRank : ordonner le web en tant que tel

Présentation, calcul et propriétés

Guillaume Seguin

26/01/2009

Plan

- 1 **Présentation du PageRank**
 - Présentation
 - Calcul

- 2 **Propriétés**
 - Balance d'énergie
 - Calculabilité et complexité
 - Calcul sur un graphe évoluant dans le temps
 - Complexité temporelle

Problématique

- Indexer des milliards de pages en constante évolution
- Introduire des critères plus forts que le contenu pour le classement des pages indexées, prenant en compte la structures topologique de l'internet
- Résister aux techniques de *web spamming*

Le PageRank

- Développé par Larry Page à l'Université de Stanford, puis avec Sergey Brin, publié (dans sa version originale) en 1998
- Utilisé comme un des facteurs intervenant dans le classement des résultats pour les recherches Google
- Se base sur des travaux d'Eugene Garfield sur l'analyse de l'importance des articles scientifiques comme fonction du nombre de citations de ces articles dans d'autres articles
- Introduit une notion d'**autorité** de chaque page web, basée sur les liens entrant et sortant de cette page web, et pas sur son contenu

The Google logo is displayed in its classic multi-colored font (blue, red, yellow, green, red) with a trademark symbol. It is centered on the white background of the slide.

Calcul

Définition : Formule de calcul du PageRank

Soit x_p le PageRank de la page p , h_p le nombre de liens sortant de cette page, I_p l'ensemble des pages pointant vers p et d un facteur d'atténuation, on a :

$$x_p = d \sum_{q \in I_p} \frac{x_q}{h_q} + (1 - d)$$

Version vectorielle

On peut étendre cette notion sous forme vectorielle :

Définition vectorielle du PageRank

Soit \vec{x} le vecteur dont les composantes sont les x_p , $W = (w_{i,j})$ la matrice de transition avec $w_{i,j} = 1/h_j$ si il existe un lien de la page j à la page i et $w_{i,j} = 0$ sinon et $\vec{\mathbb{1}}$ le vecteur dont toutes les composantes valent 1. Alors :

$$\vec{x} = dW\vec{x} + (1 - d)\vec{\mathbb{1}}$$

Algorithme simple de calcul du PageRank

Le PageRank peut être calculé par l'algorithme itératif simple suivant, qui n'est autre que la méthode de Jacobi de résolution des systèmes linéaires :

$$\vec{x}(t) = dW\vec{x}(t-1) + (1-d)\vec{\mathbb{1}}$$

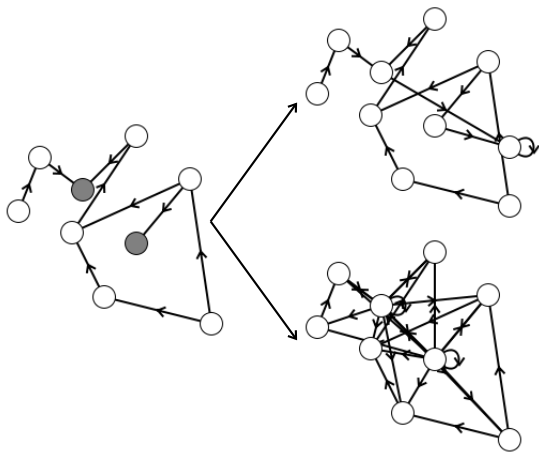
Pour $0 \leq d < 1$, ce système converge vers le vecteur \vec{x} défini précédemment.

Nous verrons plus loin la complexité en temps de cet algorithme.

Élimination des *dangling pages*

- W est presque une matrice stochastique (matrice dont la somme de chaque colonne vaut 1)
 - Posent problème les colonnes liées aux pages n'ayant pas de liens vers l'extérieur : les **dangling pages**
 - Deux solutions :
 - Ajouter une page pointant vers elle même et vers laquelle pointent toutes les pages sans liens
 - Faire pointer les pages sans liens sur toutes les pages
- Ces deux solutions impliquent donc de modifier W en \widetilde{W} , puis de ramener le vecteur \widetilde{x} calculé au vecteur x

Élimination des *dangling pages* (bis)



Importance du facteur d'amortissement

- $d = 0 \rightarrow$ toutes les pages ont un PageRank de 1
- $d = 1 \rightarrow$ le calcul peut ne pas converger, et de nombreuses pages se retrouvent avec un PageRank de 0
- Pour $d \rightarrow 1$, le PageRank des pages **inessentielles** tend vers 0 ; ces pages sont celles dont il est possible de partir sans pouvoir y revenir

Communauté

Définition : Communauté

Une **communauté** est un ensemble de pages I reliées par un quelconque lien (un site web, un ensemble de pages web de chercheurs travaillant sur un même sujet...)

On définit également :

- $out(I)$ est l'ensemble des pages extérieures à I pointées par des pages de I
- $in(I)$ est l'ensemble des pages extérieures à I pointant vers des pages de I
- $dp(I)$ est l'ensemble des pages de la communauté sans liens

Énergie

Définition : Énergie

L'**énergie** de la communauté I , notée E_I , est égale à la somme des PageRanks des pages de la communauté.

$$E_I = \sum_{p \in I} x_p$$

Similairement, on définit les énergies correspondant aux ensembles $out(I)$, $in(I)$ et $dp(I)$:

- E_I^{out} l'énergie allant de I vers l'extérieur de I
- E_I^{in} l'énergie allant de l'extérieur de I vers I
- E_I^{dp} l'énergie perdue depuis I dans les *dangling pages*

Balance d'énergie

Définition : Balance d'énergie

Ces énergies sont quantifiées par ces trois expressions qualifiées d'**équations de balance d'énergie**, avec $f_p(l)$ le ratio du nombre de pages pointées par p qui sont dans l sur le nombre total de pages pointées par p :

$$E_l^{in} = \frac{d}{1-d} \sum_{p \in in(l)} f_p(l) x_p$$

$$E_l^{out} = \frac{d}{1-d} \sum_{p \in out(l)} (1 - f_p(l)) x_p$$

$$E_l^{dp} = \frac{d}{1-d} \sum_{p \in dp(l)} x_p$$

Ces énergies sont reliées par la relation suivante, avec $|I|$ le nombre de pages de I :

Théorème

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}$$

Interactions entre communautés

Cette relation nous permet de tirer quelques conclusions sur comment optimiser les pages d'une communauté pour augmenter les scores de ses pages :

- Vu que $|I|$ est la composante de base de E_I , il vaut mieux répartir le contenu sur plusieurs petites pages que de tout rassembler
- On a $E_I \leq |I| + E_I^{in}$: les petites communautés vers lesquelles il y a peu de liens ne pourront pas avoir de hauts PageRanks

Interactions entre communautés (bis)

- Les pages sans liens sont à l'origine d'une perte d'énergie, perte d'autant plus faible que les pages pointant vers celles sans liens pointent vers beaucoup de pages
- Les liens vers l'extérieur sont à l'origine d'une perte d'énergie, d'autant plus grande si le PageRank est élevé, mais atténuée si la page pointe vers beaucoup de Pages de la communauté

Problème du *spamming*

Le **spamming** consiste à faire pointer beaucoup de pages vers une même page afin de faire monter son PageRank.
Le PageRank se comporte de manière plutôt robuste face au *spamming* : les petites communautés à faible énergie ne peuvent pas modifier significativement le PageRank.
Toutefois, il reste vulnérable aux opérations vraiment massives, qualifiées de *Google Bombing*

Calcul sur un graphe évoluant dans le temps

Théorème

Le PageRank peut être calculé en utilisant une matrice de transition $W = W(t)$ évoluant dans le temps, par exemple mise à jour au fur et à mesure de l'exploration du web par les robots indexeurs.

Complexité temporelle

Théorème

Le PageRank peut être calculé par un algorithme en temps

$$O(|H| \cdot \log(1/\epsilon))$$

Pivot de Gauss vs. algorithme de Jacobi

On note N le nombre de pages web considérées. Pour l'ordre de grandeur, en 2004, on avait $N \simeq 10^{10}$, avec une croissance exponentielle

- La méthode du pivot de Gauss nécessite $\mathcal{O}(N^3)$ opérations pour résoudre le système portant sur \vec{x}
- L'algorithme de Jacobi s'exécute en $\mathcal{O}(m \cdot |H|)$ avec $|H|$ le nombre de liens et m le nombre d'itérations
- $|H| \ll N^2$ (par nature du web)
- $m \ll N$ (expérimentalement)
- Soit $m \cdot |H| \ll N^3$: Jacobi l'emporte !

Choix pratique de l'algorithme

- D'autres algos de résolution sont en théorie plus performants (méthode de Gauss-Seidel)
- Mais problème d'implémentation : W énorme !
- Avantage de Jacobi sur les autres méthodes : peut travailler avec les données directement fournies par les robots d'indexation (pour une page donnée, la liste des liens issus de cette page)
- → Jacobi souvent choisi en pratique

Questions ?