Un vecteur propre coté en bourse : les maths de Google

Mario Lambert[1]

[1]Faculté des sciences
Université de Sherbrooke

Club mathématique – 20 septembre 2007

Plan de la présentation

Plan de la présentation

Plan de la présentation

Plan de la présentation

Plan de la présentation

Plan de la présentation

Plan de la présentation

QU'EST-CE QU'UN MOTEUR DE RECHERCHE ?

DEFINITION

Un moteur de recherche, c'est un outil qui accomplit essentiellement trois tâches :

- parcourir le Web à la recherche des pages publiques ;
- indexer ces pages de façon à pouvoir les fouiller intelligemment pour des mots-clés ou des phrases ;
- ordonner les pages par ordre d'importance

Le prof

L'homme

Jon Kleinberg (aka Rebel King)
Professeur à Cornell University
Il publie en 1999 dans le journal de l'ACM l'article *Authoritative Sources in a Hyperlinked Environment*

Son bébé

C'est l'algorithme de recherche HITS, qu'il ne commercialisera pas lui-même, mais qui se transformera en *Teoma* plus tard, qui est ce qui fait rouler *ask.com*.

Le prof

L'homme

Jon Kleinberg (aka Rebel King)
Professeur à Cornell University
Il publie en 1999 dans le journal de l'ACM l'article *Authoritative Sources in a Hyperlinked Environment*

Son bébé

C'est l'algorithme de recherche HITS, qu'il ne commercialisera pas lui-même, mais qui se transformera en *Teoma* plus tard, qui est ce qui fait rouler *ask.com*.

Les étudiants

Les hommes

Sergey Brin et Larry Page
Doctorants à Stanford University
Ils présentent en 1998 à la 7th international WWW conf.
l'article *The PageRank Citation Ranking : Bringing Order to the Web*

Leur bébé

C'est l'algorithme de recherche PageRank, qu'ils commercialiseront eux-même, et qui se transformera en *Google* plus tard.

Les étudiants

Les hommes

Sergey Brin et Larry Page
Doctorants à Stanford University
Ils présentent en 1998 à la 7th international WWW conf.
l'article *The PageRank Citation Ranking : Bringing Order to the Web*

Leur bébé

C'est l'algorithme de recherche PageRank, qu'ils commercialiseront eux-même, et qui se transformera en *Google* plus tard.

Ce qu'utilise Google

1. Indice de recherche documentaire
2. Le titre de la page Web
3. Les ancres
4. La section `<heading>`
5. Les balises `<title>`, `<alt>`, `<meta>`
6. Le nom du domaine
7. Les noms des fichiers et répertoires
8. La densité des mots-clés
9. L'importance des pages vers lesquelles on pointe
10. PageRank

Vous pouvez voir une approximation du PageRank d'une page en installant la *Google Toolbar*.

Ce qu'utilise Google

1. Indice de recherche documentaire
2. Le titre de la page Web
3. Les ancres
4. La section `<heading>`
5. Les balises `<title>, <alt>, <meta>`
6. Le nom du domaine
7. Les noms des fichiers et répertoires
8. La densité des mots-clés
9. L'importance des pages vers lesquelles on pointe
10. PageRank

Vous pouvez voir une approximation du PageRank d'une page en installant la *Google Toolbar*.

Ce qu'utilise Google

1. Indice de recherche documentaire
2. Le titre de la page Web
3. Les ancres
4. La section `<heading>`
5. Les balises `<title>, <alt>, <meta>`
6. Le nom du domaine
7. Les noms des fichiers et répertoires
8. La densité des mots-clés
9. L'importance des pages vers lesquelles on pointe
10. PageRank

Vous pouvez voir une approximation du PageRank d'une page en installant la *Google Toolbar*.

Ce qu'utilise Google

1. Indice de recherche documentaire
2. Le titre de la page Web
3. Les ancres
4. La section `<heading>`
5. Les balises `<title>, <alt>, <meta>`
6. Le nom du domaine
7. Les noms des fichiers et répertoires
8. La densité des mots-clés
9. L'importance des pages vers lesquelles on pointe
10. PageRank

Vous pouvez voir une approximation du PageRank d'une page en installant la *Google Toolbar*.

La base

- L'importance d'une page P_i est définie par

$$I(P_i) = \sum_{\substack{P_j \neq P_i \\ P_j \to P_i}} \frac{I(P_j)}{l_j}$$

où l_j est le nombre de liens sortant de la page P_j.

- La matrice de Google est $H = [H_{ij}]$ où

$$H_{ij} = \begin{cases} \frac{1}{l_j} & \text{si } P_i \to P_j \\ 0 & \text{sinon.} \end{cases}$$

La base

- L'importance d'une page P_i est définie par

$$I(P_i) = \sum_{\substack{P_j \neq P_i \\ P_j \to P_i}} \frac{I(P_j)}{l_j}$$

où l_j est le nombre de liens sortant de la page P_j.

- La matrice de Google est $H = [H_{ij}]$ où

$$H_{ij} = \begin{cases} \frac{1}{l_j} & \text{si } P_i \to P_j \\ \\ 0 & \text{sinon.} \end{cases}$$

FORMULATION EN TERME DE VECTEUR PROPRE

$$
\begin{aligned}
I(P_i) &= \sum_{\substack{P_j \neq P_i \\ P_j \to P_i}} \frac{I(P_j)}{l_j} \\
&= \sum_{j \neq i} I(P_j) H_{ji} \\
&= H_{1i} I(P_1) + H_{2i} I(P_2) + \ldots + H_{ni} I(P_n) \\
&= [H^t I]_i
\end{aligned}
$$

où I est le vecteur formé des importance de toutes les pages du Web.

Premier problème

Est-ce que H^t admet toujours 1 comme valeur propre?

Theorem

Si la somme des entrées sur chaque ligne de H est 1, alors H^t a 1 comme valeur propre.

Deuxième problème

Si 1 est une valeur propre de H^t, est-ce que le vecteur I est uniquement déterminé?

Theorem

Si $\overline{H}$ est strictement positive et que la somme des entrées sur chaque ligne de $\overline{H}$ est 1, alors l'espace propre associé à la valeur propre 1 est de dimension 1.

ALGORITHME

A une matrice carrée.
Les valeurs propres de A sont $1 = \lambda_1 > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|$.
Les vecteurs propres de A sont $x_1, x_2, \ldots x_n$.
Supposons que les vecteurs propres de A forment une base de $\mathbb{R}^n$.
Soit $I^{(0)}$ un vecteur quelconque non nul.
Alors, $I^{(0)} = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$. Posons $I^{(k)} = A I^{(k-1)}$.

THEOREM

$I^{(k)} \longrightarrow x_1$

PREMIER PROBLÈME

Est-ce que dans le cas de $\overline{\overline{H}}$,

$$1 = \lambda_1 > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|$$

VITESSE DE CONVERGENCE

THEOREM

$\lambda_2 = \alpha$

VARIATIONS SUR UN MÊME THÈME

- $\alpha = 0{,}85$
- $\alpha = 0{,}9999$
- $\alpha = 0{,}0001$

VARIATIONS SUR UN MÊME THÈME

- $\alpha = 0{,}85$
- $\alpha = 0{,}9999$
- $\alpha = 0{,}0001$

VARIATIONS SUR UN MÊME THÈME

- $\alpha = 0{,}85$
- $\alpha = 0{,}9999$
- $\alpha = 0{,}0001$

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

Les équivalences

- Web = chaîne de Markov
- H = matrice de transition
- $\overline{H}$ est stochastique
- Le vecteur I des rangs est une distribution limite
- Les composantes de I représentent la proportion du temps où un promeneur sur le Web est dans une page donnée
- connexité forte = irréductibilité de la chaîne de Markov
- Si 1 est une valeur propre multiple, il y aura une page qui va bouffer tous les rangs, et c'est un état absorbant de la chaîne de Markov.

RÉÉCRITURE DU PROBLÈME

Est-ce que ça marche?

Theorem

- $\mathbb{1} - \alpha\overline{H}$ *est non singulière*
- *La somme des éléments sur une ligne de* $\mathbb{1} - \alpha\overline{H}$ *est* $1 - \alpha$.
- $\|\mathbb{1} - \alpha\overline{H}\|_\infty = 1 + \alpha$
- $\| \left(\mathbb{1} - \alpha\overline{H}\right)^{-1} \|_\infty = \frac{1}{1-\alpha}$
- $\mathrm{cond}(\mathbb{1} - \alpha\overline{H}) = \frac{1+\alpha}{1-\alpha}$

Est-ce que ça marche?

Theorem

- $\mathbb{1} - \alpha\overline{H}$ *est non singulière*
- *La somme des éléments sur une ligne de* $\mathbb{1} - \alpha\overline{H}$ *est* $1 - \alpha$.
- $\|\mathbb{1} - \alpha\overline{H}\|_\infty = 1 + \alpha$
- $\| \left(\mathbb{1} - \alpha\overline{H}\right)^{-1} \|_\infty = \frac{1}{1-\alpha}$
- $\mathrm{cond}(\mathbb{1} - \alpha\overline{H}) = \frac{1+\alpha}{1-\alpha}$

Est-ce que ça marche?

Theorem

- $\mathbb{1} - \alpha\overline{H}$ *est non singulière*
- *La somme des éléments sur une ligne de* $\mathbb{1} - \alpha\overline{H}$ *est* $1 - \alpha$.
- $\|\mathbb{1} - \alpha\overline{H}\|_\infty = 1 + \alpha$
- $\| \left(\mathbb{1} - \alpha\overline{H}\right)^{-1} \|_\infty = \frac{1}{1-\alpha}$
- $\mathrm{cond}(\mathbb{1} - \alpha\overline{H}) = \frac{1+\alpha}{1-\alpha}$

Est-ce que ça marche?

Theorem

- $\mathbb{1} - \alpha\overline{H}$ *est non singulière*
- *La somme des éléments sur une ligne de* $\mathbb{1} - \alpha\overline{H}$ *est* $1 - \alpha$.
- $\|\mathbb{1} - \alpha\overline{H}\|_\infty = 1 + \alpha$
- $\| \left(\mathbb{1} - \alpha\overline{H}\right)^{-1} \|_\infty = \frac{1}{1-\alpha}$
- $\mathrm{cond}(\mathbb{1} - \alpha\overline{H}) = \frac{1+\alpha}{1-\alpha}$

Est-ce que ça marche?

Theorem

- $\mathbb{1} - \alpha\overline{H}$ *est non singulière*
- *La somme des éléments sur une ligne de* $\mathbb{1} - \alpha\overline{H}$ *est* $1 - \alpha$.
- $\|\mathbb{1} - \alpha\overline{H}\|_\infty = 1 + \alpha$
- $\| \left(\mathbb{1} - \alpha\overline{H}\right)^{-1} \|_\infty = \frac{1}{1-\alpha}$
- $\mathrm{cond}(\mathbb{1} - \alpha\overline{H}) = \frac{1+\alpha}{1-\alpha}$

PISTES DE SOLUTIONS

- Jacobi
- Gauss-Seidel

PISTES DE SOLUTIONS

- Jacobi
- Gauss-Seidel

QUELQUES PISTES DE SOLUTIONS

QUELQUES AVENUES INTÉRESSANTES

POUR EN SAVOIR PLUS I


Amy N. Langville et Carl D. Meyer
Deeper Inside PageRank.
Internet Mathematics Journal, vol. 1, no. 3, pp. 335-380.


Kurt Bryan et Tanya Leise
The $ 25 000 000 000 Eigenvector : The Linear Algebra Behind Google.
Rose-Hulman Institute of Tehcnology and Amherst College


David Austin
How Google Finds your Needle in the Web's Haystack.
tiré du site de l'AMS.