Comment fonctionne Google ?

Michael Eisermann

Universität Stuttgart

Conférence le 26 octobre 2009
Document mis à jour le 6 novembre 2009

Journées nationales de l'APMEP à Rouen, 24–27 octobre 2009
« Explorer les mathématiques, les mathématiques pour explorer »

`www.igt.uni-stuttgart.de/eiserm/popularisation/#google`

Introduction

Depuis plus d'une décennie, Google domine le marché des moteurs de recherche sur internet. Son point fort est qu'il trie intelligemment ses résultats par ordre de pertinence. Comment est-ce possible ?

Depuis sa conception en 1998, Google continue à évoluer et la plupart des améliorations demeurent des secrets bien gardés.

L'idée principale, par contre, a été publiée par ses fondateurs, Sergey Brin et Larry Page, dans un article de 1998 :

Le pilier de son succès est une judicieuse modélisation mathématique. L'objectif de cet exposé est de l'expliquer.

Il va sans dire que l'application pratique est devenue très importante. Bien qu'élémentaires, les arguments mathématiques sous-jacents n'en sont pas moins intéressants : l'approche fait naturellement intervenir l'algèbre linéaire, la marche aléatoire sur un graphe et le théorème du point fixe. Tout ceci en fait un très beau sujet pour la culture des mathématiques et leurs applications.

Plan de l'exposé

1. Origines et motivations
   - L'entreprise Google
   - Que fait un moteur de recherche ?
   - Structure hypertexte : le web est un graphe !
2. Comment définir la pertinence d'une page web ?
   - Premier modèle : comptage naïf
   - Second modèle : comptage pondéré
   - Troisième modèle : comptage récursif
3. Développement mathématique
   - Reformulations matricielle et probabiliste
   - Le modèle PageRank utilisé par Google
   - Le théorème du point fixe

L'entreprise Google

D'un projet d'étudiants à une entreprise mondiale :

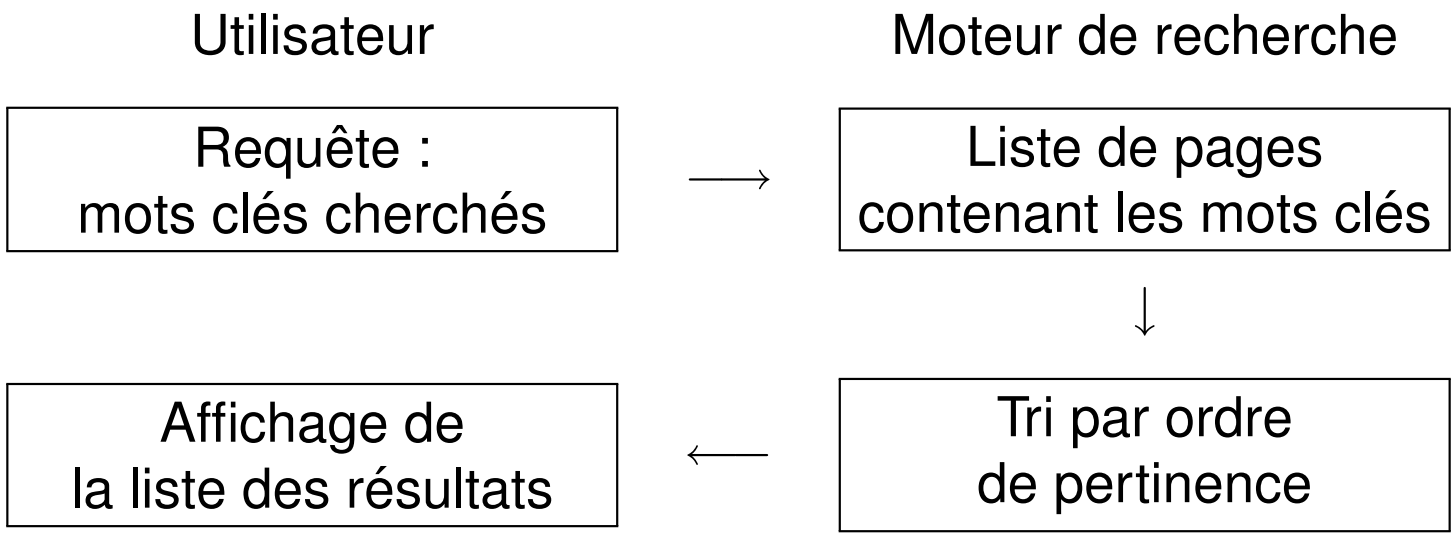
- Fondée en 1998 par Sergey Brin et Larry Page.
- Depuis 2000 vente de publicités.
- Août 2004 lancement en bourse.

Valeur en bourse	(octobre 2009)	plus de 100 milliards USD
Chiffres d'affaires	(année 2008)	21.8 milliards USD
Résultat net	(année 2008)	4.2 milliards USD
Employés	(sept. 2009)	environ 20 000
Serveurs/PC	(estimation)	plus de 500 000

📖 D. Vise, M. Malseed : *Google story*, Dunod, Paris, 2006

📖 Wikipédia, `http://fr.wikipedia.org/wiki/Google`

Que fait un moteur de recherche ?



Ingrédients cruciaux :

1. Modélisation mathématique :
   Comment définir/calculer la pertinence d'une page web ?
2. Traitement informatique :
   Comment stocker/traiter d'énormes quantités de données ?
3. Stratégie financière :
   Comment générer des bénéfices à partir d'un service gratuit ?

Objectif de cet exposé

Dans cet exposé je ne parlerai que de la partie mathématique (modélisation puis analyse) puisqu'il est destiné à un public mathématique. Les idées sont élémentaires mais puissantes.

Quand on implémente de tels modèles (même en miniature, disons dans un cours de programmation ou algorithmique mathématique) la partie informatique prend de l'ampleur (arbitrairement grande).

Il faudrait dans ce cas des techniques de tri et de recherche.
En taille moyenne des données, des méthodes standards suffisent.
Sur l'échelle mondiale, on doit bien sûr optimiser ces méthodes selon les contraintes (mémoire totale) et exigences (temps de réponse).

Finalement, si vous voulez fonder et faire tourner votre propre entreprise commerciale, s'ajoutent de profondes questions de stratégies et de financement. Je n'en parlerai pas du tout ici.

L'anarchie du web

Contenu hétérogène : toute sorte d'information,
mais peu structurée et peu hiérarchisée.

Contributions dispersées : Une multitude d'auteurs ajoutent constamment de nouvelles pages et modifient les pages existantes.

Syntaxe commune : hypertext markup language (HTML)

- Structuration logique (titres, sous-titres, paragraphes, . . .)

```
<h1> Le Titre </h1>
<p> Ceci est un paragraphe. </p>
```

- Apparence graphique (police, gras, cursif, couleur, . . .)

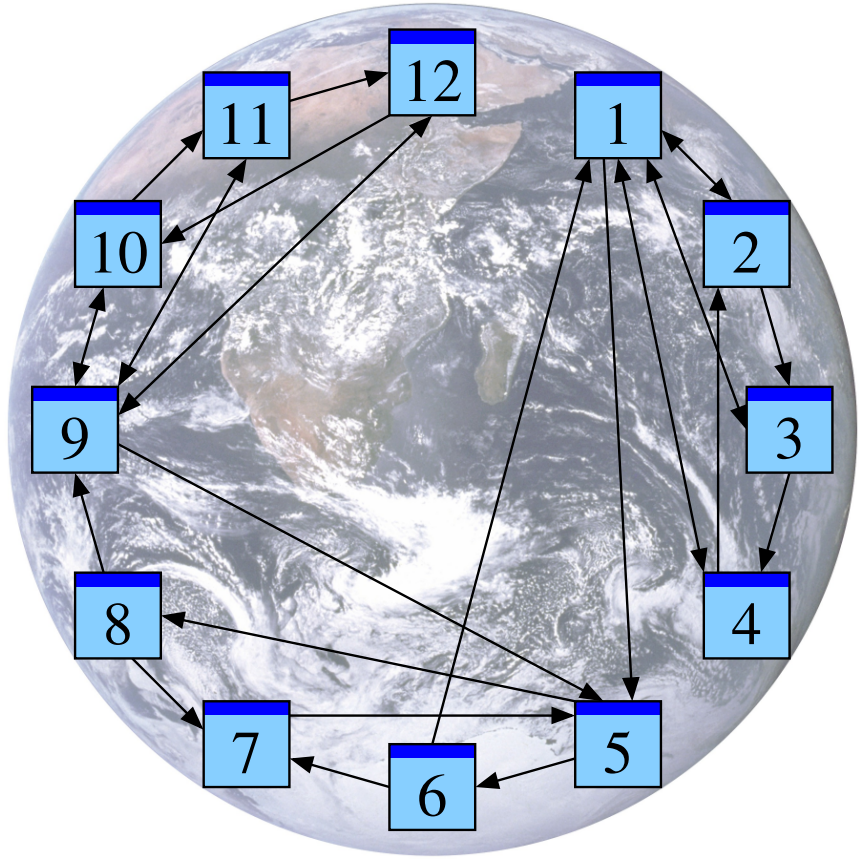
```
<b> Ceci est un texte en gras. </b>
<i> Ceci est un texte cursif. </i>
```

- Liens (références, citations entre les pages)

```
<a href="http://www.igt.uni-stuttgart.de/eiserm/">
Cliquer ici pour aller sur ma page web. </a>
```

Le web est un graphe !

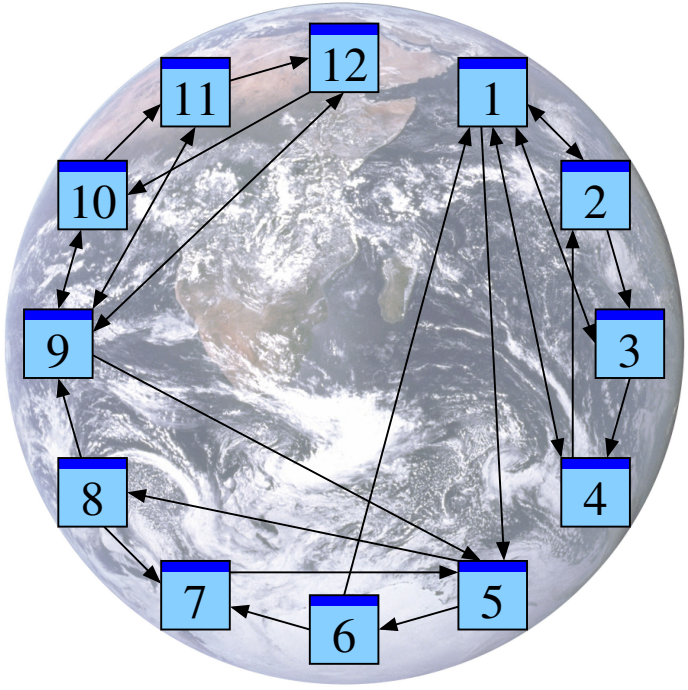
Exemple en miniature :



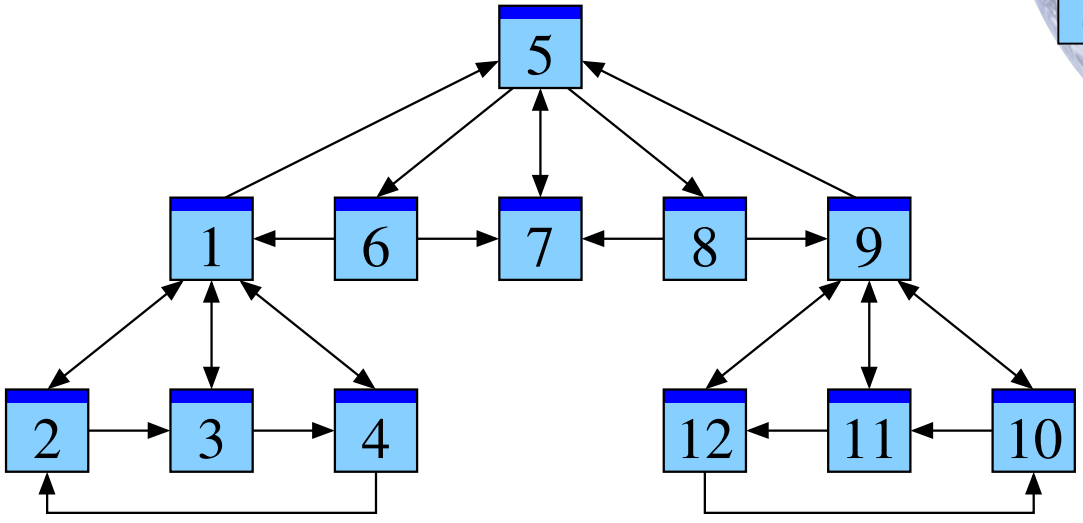
Notation : J'écris $j \to i$ pour un lien de la page P_j vers la page P_i.
Ainsi notre graphe peut s'écrire comme $1 \to 2,3,4,5$; $2 \to 1,3$;
$3 \to 1,4$; $4 \to 1,2$; $5 \to 6,8$; $6 \to 1,7$; $7 \to 5$; $8 \to 7,9$;
$9 \to 5,10,11,12$; $10 \to 9,11$; $11 \to 9,12$; $12 \to 9,10$.

Comment exploiter ce graphe ?

Comment hiérarchiser notre graphe ?



En voici une proposition ad hoc :



Comment le faire en général ? Suivant quelles heuristiques ?

Justification heuristique dans notre exemple

Nous partons de la structure brute du graphe $1 \to 2,3,4,5;$
$2 \to 1,3;$ $3 \to 1,4;$ $4 \to 1,2\,;$ $5 \to 6,8;$ $6 \to 1,7;$ $7 \to 5;$
$8 \to 7,9;$ $9 \to 5,10,11,12\,;$ $10 \to 9,11;$ $11 \to 9,12;$ $12 \to 9,10.$

L'organisation dans le plan est une information supplémentaire pour bien visualiser. Elle nécessite des choix, a priori arbitraires. J'essaie de justifier heuristiquement l'hiérarchie proposée ci-dessus :

Parmi les pages P_1, P_2, P_3, P_4 la page P_1 sert de référence commune et semble un bon point de départ pour chercher des informations.

Il en est de même dans le groupe $P_9, P_{10}, P_{11}, P_{12}$
où la page P_9 sert de référence commune.

La structure de P_5, P_6, P_7, P_8 est similaire, où P_7 est la plus citée.

À noter toutefois que les pages P_1 et P_9, déjà reconnues comme importantes, font référence à la page P_5.

On pourrait ainsi soupçonner que la page P_5 contient de l'information essentielle pour l'ensemble, qu'elle est la plus pertinente.

Premier modèle : comptage naïf

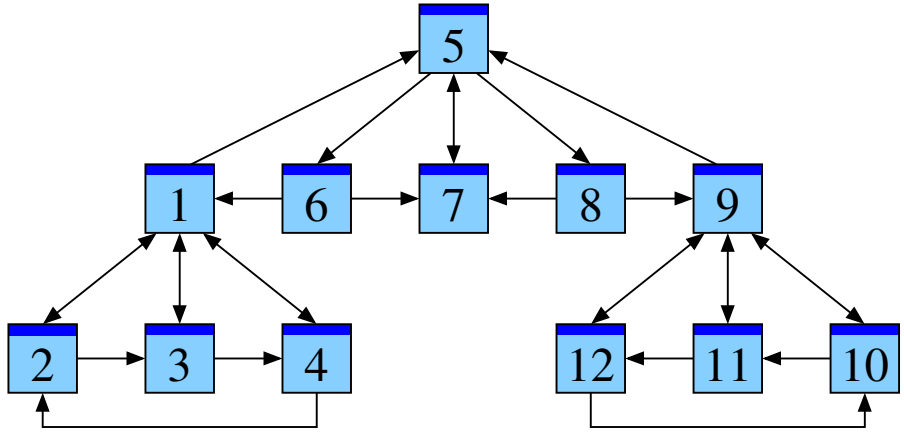
Heuristique : *Une page importante reçoit beaucoup de liens.*

Avec un peu de naïveté on croira aussi la réciproque :
Si une page reçoit beaucoup de liens, alors elle est importante.

Première tentative d'une définition mathématique :

$$m_i := \sum_{j \to i} 1.$$

Appliquons-la à notre exemple :



Ici on trouve $m_1 = m_9 = 4$ devant $m_5 = m_7 = 3$.

Discussion du premier modèle

Dans ce premier modèle on définit l'importance m_i de la page P_i comme le nombre des liens $j \to i$ reçus par P_i.

Autrement dit, m_i est égal au nombre de « votes »
pour la page P_i, où chaque vote contribue par la même valeur 1.

Avantage : C'est facile à définir et à calculer.

Inconvénient : Le résultat ne correspond souvent pas à l'importance ressentie par l'utilisateur. Dans notre exemple on trouve $m_1 = m_9 = 4$ devant $m_5 = m_7 = 3$, contrairement à notre classement ad hoc.

Manipulabilité : Ce qui est pire, ce comptage naïf est trop facile à manipuler en ajoutant des pages sans intérêt recommandant une page quelconque.

Second modèle : comptage pondéré

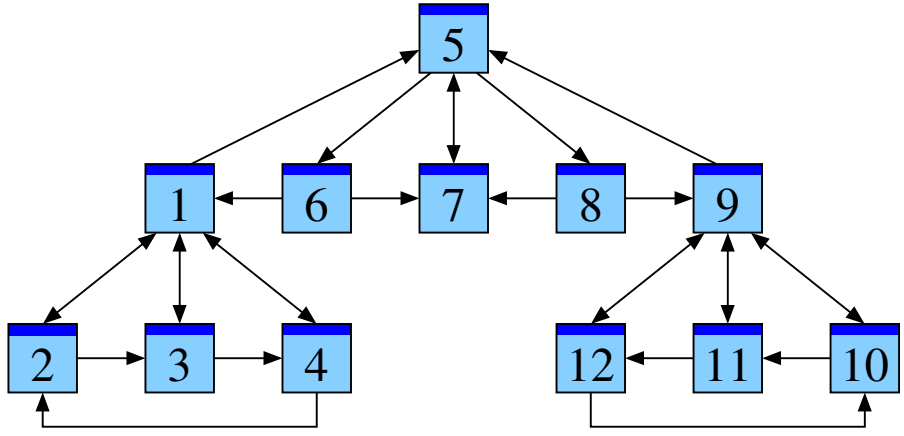
Heuristique : *Un lien $j \to i$ est un vote de la page P_j en faveur de P_i.*

Supposons d'abord que toutes les pages ont un poids égal.
Nous partageons le vote de la page P_j en ℓ_j parts égales.

Seconde tentative d'une définition mathématique :

$$m_i := \sum_{j \to i} \frac{1}{\ell_j}.$$

Appliquons-la à notre exemple :



Ici on trouve $m_1 = m_9 = 2$ devant $m_5 = 3/2$ et $m_7 = 4/3$.

Discussion du second modèle

Certaines pages émettent beaucoup de liens :
ceux-ci semblent moins spécifiques et leur poids sera plus faible.

Nous partageons donc le vote de la page P_j en ℓ_j parts égales,
où ℓ_j dénote le nombre de liens émis.

Autrement dit, dans ce second modèle la valeur m_i compte
le nombre de « votes pondérés » pour la page P_i.

Avantage : C'est facile à définir et à calculer.

Inconvénient : Le résultat ne correspond toujours pas bien
à l'importance ressentie par l'utilisateur.

Manipulabilité : Comme avant ce comptage est trop facile
à truquer en ajoutant des pages artificielles.

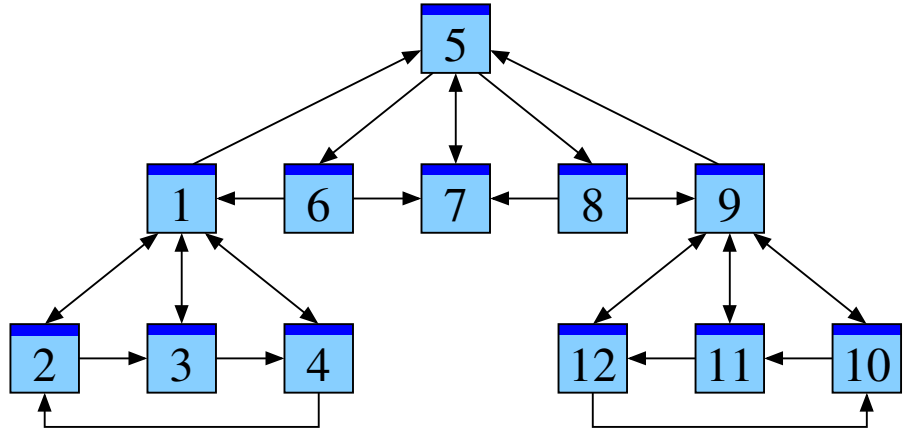
Troisième modèle : comptage récursif

Heuristique :
Une page est importante si beaucoup de pages importantes la citent.

Troisième tentative d'une définition mathématique :

$$m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j.$$

Appliquons-la à notre exemple :



Ici on trouve

$$m = (\ \underset{P_1}{2},\ \underset{P_2}{1},\ \underset{P_3}{1},\ \underset{P_4}{1},\ \underset{P_5}{3},\ \underset{P_6}{1},\ \underset{P_7}{2},\ \underset{P_8}{1},\ \underset{P_9}{2},\ \underset{P_{10}}{1},\ \underset{P_{11}}{1},\ \underset{P_{12}}{1}\ ).$$

Discussion du troisième modèle

Dans ce troisième modèle, le poids du vote $j \to i$
est proportionnel au poids m_j de la page émettrice.

Avantage : Contrairement aux modèles précédents,
la page P_5 est repérée comme la plus importante.
C'est bon signe, nous sommes sur la bonne piste...

Le modèle est facile à formuler mais moins évident à calculer !

Calcul : Pour trouver une solution, remarquons qu'il s'agit
d'un système de n équations linéaires à n inconnues.
J'ai fait ce calcul pour vous... Disposant du résultat énoncé ici,
vous pouvez facilement vérifier que c'est effectivement une solution.
(C'est en fait la seule, à multiplication par une constante près.)

Grandeur : Dans notre exemple, où $n = 12$, bien que pénible, ce
calcul est faisable à la main. C'est même très facile sur ordinateur.
Pour les graphes vraiment grands, par contre, nous aurons besoin de
méthodes spécialisées plus performantes dont on parlera plus bas.

Reformulation matricielle

Nous avons dégagé un système d'équations linéaires :

$$m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j.$$

Définissons alors la matrice $A = (a_{ij})$ par

$$a_{ij} := \begin{cases} \frac{1}{\ell_j} & \text{si } j \to i, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi notre équation s'écrit

$$Am = m$$

ou encore

$$(A - \mathrm{Id})m = 0.$$

Vive l'algèbre linéaire !

Commentaires

Notre équation peut s'écrire sous forme matricielle :
ainsi s'offrent à nous tous les outils de l'algèbre linéaire !

Sous la forme $Am = m$ c'est une équation de point fixe,
ou bien la définition d'un vecteur propre de valeur propre 1.

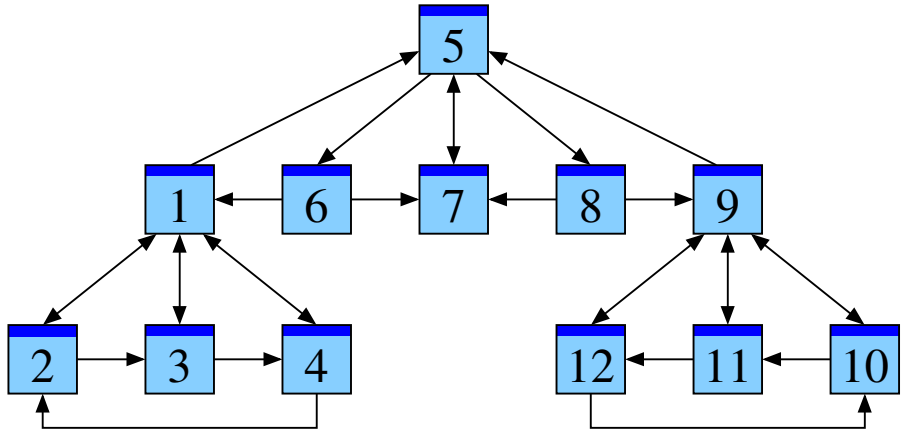
Sous la forme $(A - \mathrm{Id})m = 0$ c'est un honnête
système d'équations linéaires homogènes.

Ceci provoque des questions naturelles :

1. Existe-t-il toujours une solution à notre équation ?
2. Y en a-t-il plusieurs ? Ou une seule ?
3. Comment la calculer ? Efficacement ?

Reformulation matricielle de notre exemple

Notre graphe :



Et sa matrice :

$$A = \begin{pmatrix}
0 & 1/2 & 1/2 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/4 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/4 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/4 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/4 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/2 & 0
\end{pmatrix}.$$

L'équation $Am = m$ admet comme solution

$$m = (\ 2,\ 1,\ 1,\ 1,\ 3,\ 1,\ 2,\ 1,\ 2,\ 1,\ 1,\ 1\ )^t.$$

Commentaires

Notre construction de la matrice A assure deux propriétés :

- Tous les coefficients sont positifs ou nuls.
- La somme des coefficients dans chaque colonne vaut 1.

Une telle matrice est appelée *matrice stochastique*.

La somme des coefficients d'une ligne n'est pas constante.
(C'est le comptage pondéré des liens de notre second modèle.)

On constate aussi que notre matrice est *creuse* :
elle contient beaucoup de coefficients nuls !

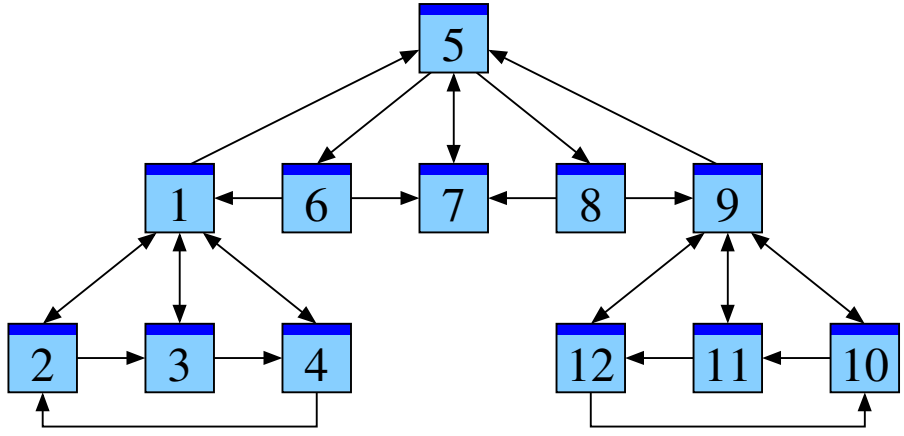
Ceci est encore plus frappant pour l'internet en taille réelle :
Il existe quelques milliards de pages web mais
typiquement chacune n'émet que très peu de liens.

Pour des applications réalistes à l'échelle nature, c'est une observation cruciale pour l'implémentation efficace des calculs !

Promenade aléatoire sur le web

Imaginons un « surfeur aléatoire » qui se balade sur internet en cliquant sur les liens au hasard. Comment évolue sa position ?

Notre graphe :



Évolution des probabilités en partant de la page P_7 :

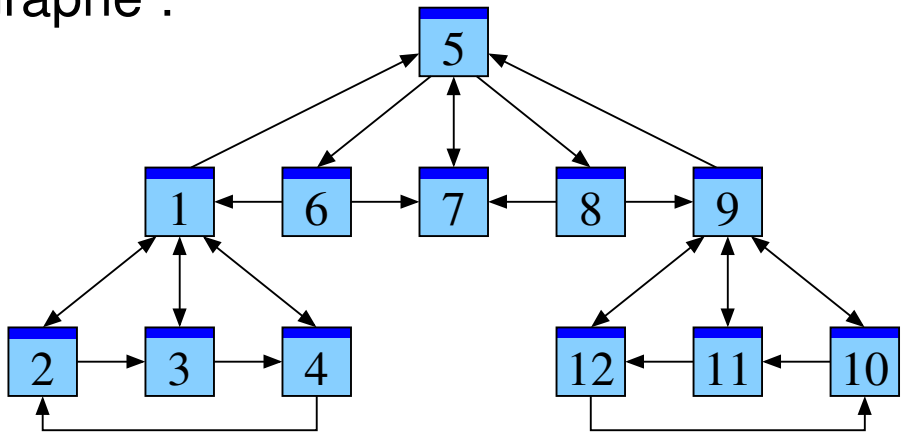
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
t=0	.000	.000	.000	.000	.000	.000	1.00	.000	.000	.000	.000	.000
t=1	.000	.000	.000	.000	1.00	.000	.000	.000	.000	.000	.000	.000
t=2	.000	.000	.000	.000	.000	.333	.333	.333	.000	.000	.000	.000
t=3	.167	.000	.000	.000	.333	.000	.333	.000	.167	.000	.000	.000
t=4	.000	.042	.042	.042	.417	.111	.111	.111	.000	.042	.042	.042
t=5	.118	.021	.021	.021	.111	.139	.250	.139	.118	.021	.021	.021
...												
t=29	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059
t=30	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059

Cette diffusion converge vers une distribution stationnaire !

Promenade aléatoire sur le web

Vérifions notre observation par un second exemple, partant de P_1.

Toujours notre graphe :



Évolution des probabilités en partant de la page P_1 :

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
t=0	1.00	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
t=1	.000	.250	.250	.250	.250	.000	.000	.000	.000	.000	.000	.000
t=2	.375	.125	.125	.125	.000	.083	.083	.083	.000	.000	.000	.000
t=3	.229	.156	.156	.156	.177	.000	.083	.000	.042	.000	.000	.000
t=4	.234	.135	.135	.135	.151	.059	.059	.059	.000	.010	.010	.010
t=5	.233	.126	.126	.126	.118	.050	.109	.050	.045	.005	.005	.005
...												
t=69	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059
t=70	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059

La mesure stationnaire est la même !

La loi de transition

Comment formaliser cette diffusion des probabilités ?

Au temps t notre surfeur se trouve sur la page P_j avec probabilité p_j.
La probabilité de partir de P_j et de suivre le lien $j \to i$ est alors $\frac{1}{\ell_j} p_j$.
La probabilité d'arriver au temps $t+1$ sur la page P_i est donc

$$p'_i := \sum_{j \to i} \frac{1}{\ell_j} p_j.$$

Cette loi de transition définit la distribution suivante, notée $p' = T(p)$.
Une mesure stationnaire est caractérisée par l'équation d'équilibre

$$m = T(m) \quad \text{c'est-à-dire} \quad m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j.$$

Vive la théorie des probabilités !

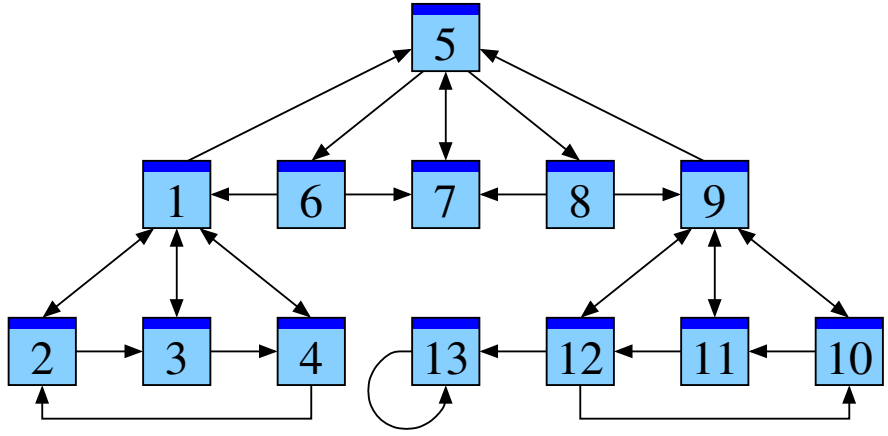
Robustesse du modèle récursif

Que se passe-t-il si l'on ajoute des pages artificielles,
sans intérêt, qui recommandent une page afin de la booster ?

Dans la mesure équilibre ces pages ont chacune la mesure 0.
À ceci près, les équilibres avant et après extension sont les mêmes.

Attention aux trous noirs !

Que se passe-t-il quand notre graphe contient une page sans issue ?



Notre surfeur aléatoire tombera tôt ou tard sur la page P_{13},
où il demeure pour le reste de sa vie.

La seule mesure stationnaire est

$$m = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1).$$

Dans ce cas notre modèle n'est pas très réaliste !

Le modèle PageRank utilisé par Google

Pour échapper aux trous noirs, Google utilise un modèle plus raffiné :

1. Téléportation :

   Avec une probabilité fixée $c \in [0, 1]$ le surfeur abandonne sa page actuelle P_j et recommence sur une des n pages du web.

2. Promenade aléatoire :

   Sinon, avec probabilité $1 - c$, le surfeur suit un des liens de la page P_j, choisi de manière équiprobable (comme avant).

Dans ce modèle la transition est donnée par

$$p'_i := \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} p_j.$$

La mesure d'équilibre vérifie donc

$$m_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} m_j.$$

Interprétation de la constante c

Deux cas extrêmes :

- Pour $c = 0$ nous obtenons le modèle précédent :
  c'est la marche aléatoire sans téléportation.
- Pour $c = 1$ le surfeur aléatoire ne suit jamais de lien :
  il saute de page en page de manière aléatoire.

Un bon choix de c se situe donc quelque part entre 0 et 1.
En général, on choisira la constante c positive mais proche de zéro.

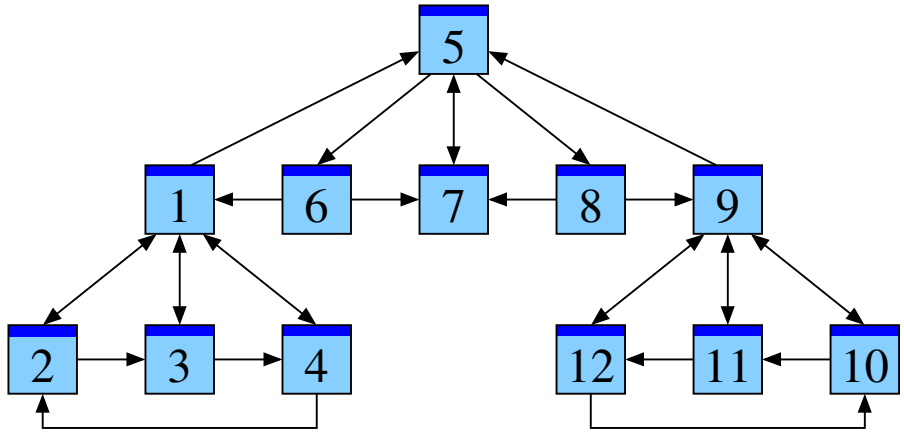
Remarque (épreuve de Bernoulli, loi binomiale)

La valeur $\frac{1}{c}$ est le *nombre moyen* de pages visitées (nombre de liens suivis plus 1) avant de recommencer sur une page aléatoire.

Par exemple, $c = 0.15$ correspond à suivre environ 6 liens en moyenne. On pourrait argumenter que ceci correspond au comportement des utilisateurs. . . à calibrer expérimentalement !

Application à notre exemple

Notre graphe :



Évolution des probabilités en partant de la page P_1, avec $c = 0.15$:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
t=0	1.00	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
t=1	.013	.225	.225	.225	.225	.013	.013	.013	.013	.013	.013	.013
t=2	.305	.111	.111	.111	.028	.076	.087	.076	.034	.020	.020	.020
t=3	.186	.124	.124	.124	.158	.021	.085	.021	.071	.028	.028	.028
t=4	.180	.105	.105	.105	.140	.057	.075	.057	.057	.040	.040	.040
t=5	.171	.095	.095	.095	.126	.052	.101	.052	.087	.042	.042	.042
...												
t=29	.120	.066	.066	.066	.150	.055	.102	.055	.120	.066	.066	.066
t=30	.120	.066	.066	.066	.150	.055	.102	.055	.120	.066	.066	.066

Robustesse du modèle PageRank

Que se passe-t-il si l'on ajoute une page sans issue ?

Dans la mesure équilibre cette page absorbe un peu des probabilités, mais pas trop grâce à la téléportation.

Que se passe-t-il si l'on ajoute des pages artificielles,
sans intérêt, qui recommandent une page afin de la booster ?

Dans la mesure équilibre ces pages ont un poids $\frac{c}{n}$, assez petit.
Ainsi les équilibres avant et après extension sont très proches.

Quel rôle jouent les mathématiques dans tout cela ?

Question initiale : Comment définir la « pertinence » des pages web ?

C'est d'abord un défi de modélisation :

- Quelles sont les données initiales ? Où veut-on aboutir ?
- Les maths fournissent un langage pour formuler nos modèles.
- Des calculs permettent de tester puis de raffiner nos modèles.

Une fois le modèle fixé, les maths permettent de l'analyser :

1. Existe-t-il toujours une solution à notre équation ?
2. Y en a-t-il plusieurs ? Ou une seule ?
3. Comment la calculer ? Efficacement ?

Le théorème du point fixe

Rappel : Dans notre modèle la loi de transition est donnée par

$$p'_i := \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} p_j. \tag{1}$$

La mesure d'équilibre vérifie donc

$$m_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} m_j. \tag{2}$$

Théorème ($\Leftarrow$ Théorème de point fixe de Banach)

Considérons un graphe fini et fixons le paramètre $c \in]0,1]$. Alors :

1. *L'équation* (2) *admet une unique solution.*
   Elle vérifie $m_1, \ldots, m_n > 0$ et $m_1 + \cdots + m_n = 1$.
2. *Pour toute distribution de probabilité initiale le processus de diffusion* (1) *converge vers cette unique mesure stationnaire m.*
3. *La convergence est au moins aussi rapide que celle de la suite géométrique $(1-c)^n$ vers 0.*

Vive l'analyse !

Commentaires

Le théorème présenté ici est une application directe
du théorème de point fixe de Banach (qui date de 1922).

Avec cet outil, la démonstration est en fait relativement simple :
on vérifie que la loi de transition $T\colon \mathbb{R}^n \to \mathbb{R}^n$ est une application
contractante de rapport $(1-c)$ dans la norme $\|\cdot\|_1$.

Le théorème sur les matrices stochastiques tel qu'il est formulé ici
est aussi un cas particulier du théorème de Perron–Frobenius :

Théorème ($\Leftarrow$ Théorème de Perron–Frobenius)

Si une matrice stochastique n'a que des coefficients positifs, alors 1
est la valeur propre de plus grande valeur absolue et l'espace propre
associé est de dimension 1*. Il admet pour base un unique vecteur*
propre de coefficients positifs de somme 1*.*

Conclusion

Pour être utile, un moteur de recherche doit non seulement *énumérer* les résultats d'une requête mais les *classer* par ordre de pertinence.

- En première approximation Google analyse le graphe formé par les pages web et les liens entre elles.
- Interprétant un lien $j \to i$ comme « vote » de la page P_j en faveur de P_i, le modèle PageRank définit une mesure de « popularité ».
- Le théorème du point fixe assure que cette équation admet une unique solution, et justifie l'algorithme itératif pour l'approcher.

Muni de ces outils mathématiques et d'une habile stratégie d'entreprise, Google gagne des milliards de dollars.

Il fallait y penser.

Je vous remercie de votre attention !

Littérature

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

📖 K. Bryan, T. Leise : *The $825,000,000,000 eigenvector : the linear algebra behind Google*, SIAM Review 48 (2006) 569-581

(Ces articles sont disponibles en ligne, cherchez-les avec Google. ;-)

📖 M. Eisermann : *Comment Google classe les pages web*, Images des Mathématiques, La Tangente, Quadrature
`www.igt.uni-stuttgart.de/eiserm/popularisation/#google`

(Expérience pratique : si vous pensez que ce document le mérite, faites-y pointer vos liens pour augmenter son PageRank. ;-)

Le modèle PageRank est-il plausible ?

La structure caractéristique des documents hypertextes sont les citations mutuelles.

- L'hypothèse à la base du modèle PageRank est que l'auteur d'une page ajoute des liens vers les pages qu'il considère utiles.
- Ainsi des millions d'auteurs lisent et jugent mutuellement leurs pages, et leurs jugements s'expriment par leurs liens.
- Le modèle de la marche aléatoire en profite en transformant l'évaluation mutuelle en une mesure globale de popularité.

Cet argument de plausibilité est à débattre et à expérimenter. . .

L'ultime argument en faveur du modèle PageRank est son succès : le classement semble bien refléter les attentes des utilisateurs.

Le modèle PageRank est-il descriptif ou normatif ?

Au début de son existence, Google se voulait un outil *descriptif* :
si une page est importante, alors elle figure en tête du classement.

Son écrasant succès a fait de Google une référence *normative* :
si une page figure en tête du classement, alors elle est importante.

Pour des sites web commerciaux, l'optimisation de leur classement PageRank est ainsi devenue un enjeu vital.

Stratégie évidente : il suffit d'attirer des liens, de préférence ceux émis des pages importantes, et il vaut mieux en émettre très peu.

Ainsi l'omniprésence de Google change l'utilisation des liens par les auteurs des pages web. . . ce qui remet en question l'hypothèse à la base même du modèle PageRank.