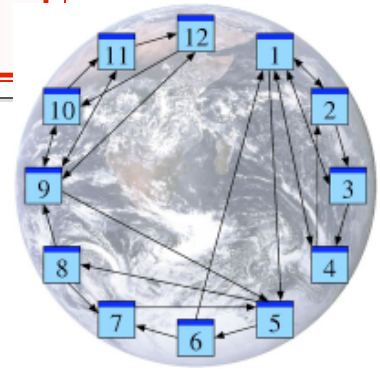


Comment Google classe les pages web

Une promenade sur la toile

Le 22 septembre 2009, par **Michael Eisermann**

Professeur à l'Université de Stuttgart en Allemagne ([page web](#))



DEPUIS 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 [1] : le pilier de son succès est une judicieuse modélisation mathématique que nous retraçons ici.

Que fait un moteur de recherche ?

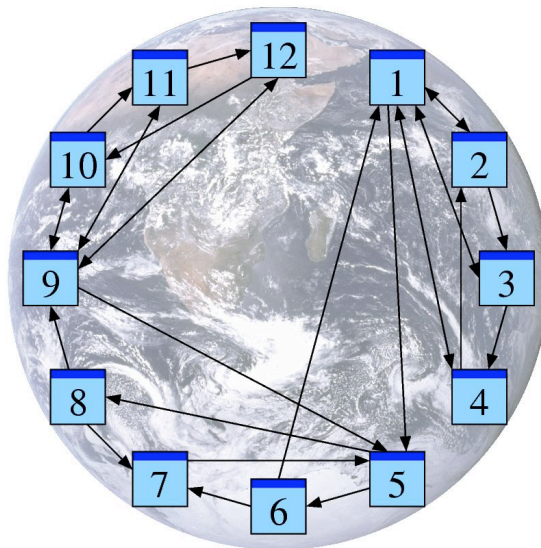
Une base de données a une structure prédéfinie qui permet d'en extraire des informations, par exemple « nom, rue, code postal, téléphone, ... ». L'internet, par contre, est peu structuré : c'est une immense collection de textes de nature variée. (Je fais abstraction ici de différents formats et médias, car pour la recherche on revient bien aux mots-clés.) Toute tentative de classification semble vouée à l'échec, d'autant plus que le web évolue rapidement : une multitude d'auteurs indépendants ajoute constamment de nouvelles pages et modifie les pages existantes.

Pour trouver une information dans ce tas amorphe, l'utilisateur pourra lancer une recherche de mots-clés. Ceci nécessite une certaine préparation pour être efficace : le moteur de recherche copie préalablement les pages web une par une en mémoire locale et trie les mots par ordre alphabétique. Le résultat est un annuaire de mots-clés avec leurs pages web associées.

Pour un mot-clé donné il y a typiquement des milliers de pages correspondantes (plus d'un million pour « tangente », par exemple). Certaines pages sont pourtant plus pertinentes que d'autres. Comment aider l'utilisateur à repérer les résultats potentiellement intéressants ? C'est ici que Google a apporté sa grande innovation.

Le web est un graphe !

Profitons du peu de structure qui soit disponible. L'internet n'est pas une collection de textes indépendants mais un immense *hypertexte* : les pages se citent mutuellement. Afin d'analyser cette structure nous allons négliger le contenu des pages et ne tenir compte que des liens entre elles. Ce que nous obtenons est la structure d'un graphe. La figure suivante montre un exemple en miniature.



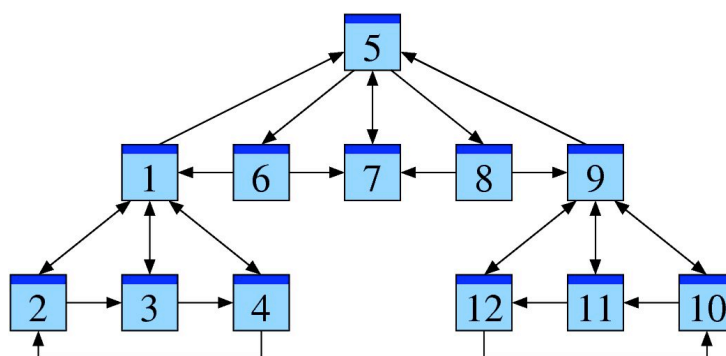
Ici les sommets du graphe représentent les pages web et les flèches représentent les liens, c'est-à-dire les citations entre pages web. Chaque flèche est orientée de la page émettrice vers la page citée. [2]

Dans la suite je note les pages web par $P_1, P_2, P_3, \dots, P_n$ et j'écris $j \rightarrow i$ si la page P_j cite la page P_i . Dans notre graphe nous avons un lien $1 \rightarrow 5$, par exemple, mais pas de lien $5 \rightarrow 1$. Néanmoins, dans ce premier exemple, toutes les pages communiquent via des chemins à un ou plusieurs pas.

Comment exploiter ce graphe ?

Les liens sur internet ne sont pas aléatoires mais ont été édités avec soin. Quels renseignements pourrait nous donner ce graphe? L'idée de base, encore à formaliser, est qu'un lien $j \rightarrow i$ est une recommandation de la page P_j d'aller lire la page P_i . C'est ainsi un vote de P_j en faveur de l'autorité de la page P_i .

Analysons notre exemple sous cet aspect. La présentation suivante de notre graphe suggère une hiérarchie possible — encore à justifier.



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 du groupe 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. Dans la suite nous allons essayer de formaliser ce classement.

Premier modèle : comptage naïf

Il est plausible qu'une page importante reçoit beaucoup de liens. Avec un peu de naïveté, on croira aussi l'affirmation réciproque : si une page reçoit beaucoup de liens, alors elle est importante. Ainsi on pourrait définir la mesure d'importance m_i de la page P_i comme le nombre des liens $j \rightarrow i$ reçus par P_i . En formule ceci s'écrit comme suit :

$$m_i := \sum_{j \rightarrow i} 1. \quad (1)$$

Ici le signe « $\sum_{j \rightarrow i}$ » dénote la somme sur tous les liens pointant vers la page P_i , et les termes à sommer valent tous 1. Autrement dit, m_i est égal au nombre de « votes » pour la page P_i , où chaque vote contribue par la même valeur 1. C'est facile à définir et à calculer, mais ne correspond souvent pas à l'importance ressentie par l'utilisateur : dans notre exemple on trouve $m_1 = m_9 = 4$ devant $m_5 = m_7 = 3$. 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

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. Ainsi on pourrait définir une mesure plus fine :

$$m_i := \sum_{j \rightarrow i} \frac{1}{\ell_j}. \quad (2)$$

Autrement dit, m_i compte le nombre de « votes pondérés » pour la page P_i . C'est facile à définir et à calculer, mais ne correspond toujours pas bien à l'importance ressentie : dans notre exemple on trouve $m_1 = m_9 = 2$ devant $m_5 = \frac{3}{2}$ et $m_7 = \frac{4}{3}$. Et comme avant ce comptage est trop facile à truquer.

Troisième modèle : comptage récursif

Heuristiquement, une page P_i paraît importante si beaucoup de pages importantes la citent. Ceci nous mène à définir l'importance m_i de manière récursive comme suit :

$$m_i = \sum_{j \rightarrow i} \frac{1}{\ell_j} m_j. \quad (3)$$

Ici le poids du vote $j \rightarrow i$ est proportionnel au poids m_j de la page émettrice. C'est facile à formuler mais moins évident à calculer... Une méthode efficace sera expliquée dans la suite [3]. Pour vous rassurer vous pouvez déjà vérifier que notre exemple admet bien la solution

$$m = \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 & P_9 & P_{10} & P_{11} & P_{12} \\ 2, & 1, & 1, & 1, & 3, & 1, & 2, & 1, & 2, & 1, & 1, & 1 \end{pmatrix}.$$

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

Remarquons que (3) est un système de n équations linéaires à n inconnues. Dans notre exemple, où $n = 12$, il est déjà pénible à résoudre à la main, mais encore facile sur ordinateur. Pour les graphes beaucoup plus grands nous aurons besoin de méthodes spécialisées.

Promenade aléatoire sur la toile

Avant de tenter de résoudre l'équation (3), essayons d'en développer une intuition. Pour ceci imaginons un surfeur aléatoire qui se balade sur internet en cliquant sur les liens au hasard. Comment évolue sa position ?

À titre d'exemple, supposons que notre surfeur démarre au temps $t = 0$ sur la page P_7 . Le seul lien pointe vers P_5 , donc au temps $t = 1$ le surfeur s'y retrouve avec probabilité 1. D'ici partent trois liens, donc au temps $t = 2$ il se trouve sur une des pages P_6, P_7, P_8 avec probabilité $\frac{1}{3}$. Voici les probabilités suivantes (arrondies à 10^{-3} près) :

	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

On observe une diffusion qui converge assez rapidement vers une distribution stationnaire (à 10^{-3} près au bout d'une trentaine d'itérations). Vérifions cette observation par un second exemple, partant cette fois-ci 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

Bien que la diffusion mette plus de temps à se stabiliser, la mesure stationnaire est la même ! Elle coïncide d'ailleurs avec notre solution $m = (2, 1, 1, 1, 3, 1, 2, 1, 2, 1, 1, 1)$, ici divisée par 17 pour normaliser la somme à 1. Les pages où m_i est grand sont les plus « fréquentées » ou les plus

« populaires ». Dans la quête de classer les pages web par ordre d'importance c'est encore un argument pour utiliser la mesure m comme indicateur.

Le modèle du surfeur aléatoire peut sembler étonnant, mais en absence d'information plus précise, le recours aux considérations probabilistes se révèle souvent très utile !

La loi de transition

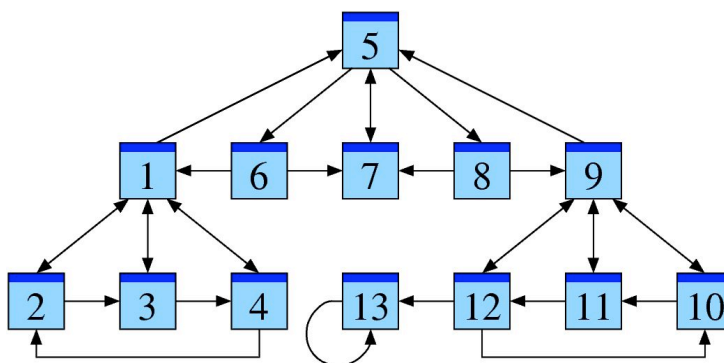
Comment formaliser la diffusion illustrée ci-dessus ? Supposons qu'au temps t notre surfeur aléatoire se trouve sur la page P_j avec une probabilité p_j . La probabilité de partir de P_j et de suivre le lien $j \rightarrow i$ est alors $\frac{1}{\ell_j} p_j$. La probabilité d'arriver au temps $t + 1$ sur la page P_i est donc

$$p'_i := \sum_{j \rightarrow i} \frac{1}{\ell_j} p_j. \quad (4)$$

Étant donnée la distribution initiale p , la loi de transition (4) définit la distribution suivante $p' = T(p)$. C'est ainsi que l'on obtient la ligne $t + 1$ à partir de la ligne t dans nos exemples. (En théorie des probabilités ceci s'appelle une *chaîne de Markov*.) La mesure stationnaire est caractérisée par l'équation d'équilibre $m = T(m)$, qui est justement notre équation (3) de départ.

Attention aux trous noirs

Que se passe-t-il quand notre graphe contient une page (ou un groupe de pages) sans issue ? Pour illustration, voici notre graphe augmenté d'une nouvelle page P_{13} sans issue :



L'interprétation comme marche aléatoire permet de résoudre l'équation (3) sans aucun calcul : la page P_{13} absorbe toute la probabilité car notre surfeur aléatoire tombera tôt ou tard sur cette page, où il demeure pour le reste de sa vie. Ainsi la solution est

$$m = (\begin{matrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 & P_9 & P_{10} & P_{11} & P_{12} & P_{13} \\ 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 1 \end{matrix}).$$

Notre modèle n'est donc pas encore satisfaisant.

Le modèle PageRank utilisé par Google

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

- avec une probabilité fixée c le surfeur abandonne sa page actuelle P_j et recommence sur une des n pages du web, choisie de manière équiprobable ;
- sinon, avec probabilité $1 - c$, le surfeur suit un des liens de la page P_j , choisi de manière équiprobable. (C'est la marche aléatoire usuelle).

Cette astuce de « téléportation » évite de se faire piéger par une page sans issue, et garantit d'arriver n'importe où dans le graphe, indépendamment des questions de connexité.

Dans ce modèle la transition est donnée par

$$p'_i := \frac{c}{n} + \sum_{j \rightarrow i} \frac{1-c}{\ell_j} p_j. \quad (5)$$

Le premier terme $\frac{c}{n}$ provient de la téléportation, le second terme est la marche aléatoire précédente. La mesure d'équilibre satisfait donc à l'équation

$$m_i = \frac{c}{n} + \sum_{j \rightarrow i} \frac{1-c}{\ell_j} m_j. \quad (6)$$

Le paramètre c est encore à calibrer. Pour $c = 0$ nous obtenons le modèle précédent (voir (3) et (4) ci-dessus). Pour $0 < c \leq 1$ la valeur $\frac{1}{c}$ est le nombre moyen de pages visitées, c'est-à-dire le nombre de liens suivis plus un, avant de recommencer sur une page aléatoire (processus de Bernoulli).

En général, on choisira la constante c non nulle mais proche de zéro. Par exemple, le choix $c = 0.15$ correspond à suivre environ 6 liens en moyenne, ce qui semble une description réaliste.

Pour conclure l'analyse de notre exemple, voici la marche aléatoire 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
$t = 1$.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

La mesure stationnaire est vite atteinte, et la page P_5 arrive en tête avec $m_5 = 0.15$ avant les pages P_1 et P_9 avec $m_1 = m_9 = 0.12$.

Le théorème du point fixe

Afin de développer un modèle prometteur nous avons utilisé des arguments heuristiques et des illustrations expérimentales. Fixons maintenant ce modèle et posons-le sur un solide fondement théorique. Nos calculs aboutissent bel et bien dans notre exemple miniature, mais est-ce toujours le cas ? Le beau résultat suivant y répond en toute généralité [4] :

Théorème du point fixe. — *Considérons un graphe fini quelconque et fixons le paramètre c tel que $0 < c \leq 1$. Alors :*

- *L'équation (6) admet une unique solution vérifiant $m_1 + \dots + m_n = 1$. Dans cette solution m_1, \dots, m_n sont tous strictement positifs.*
- *Pour toute distribution de probabilité initiale sur le graphe, le processus de diffusion (5) converge vers cette unique mesure stationnaire m .*
- *La convergence est au moins aussi rapide que celle de la suite géométrique $(1 - c)^n$ vers 0.*

Soulignons l'importance de chacun de ces trois points. Le premier assure tout simplement l'existence et l'unicité d'une solution à notre problème. Mieux encore, non seulement une solution existe mais le deuxième point nous dit comment la calculer : par un algorithme itératif. Ici l'indépendance du point de départ garantit une certaine stabilité numérique : lors des calculs avec des nombres à virgule, des erreurs d'arrondi sont souvent inévitables, mais heureusement ici de telles perturbations n'influencent pas le résultat final. Enfin, le troisième point garantit que la vitesse de convergence est suffisamment grande, ce qui est cruciale pour toute application de grande nature. Pour son classement Google traite plusieurs milliards de pages web [5]. Cette tâche herculéenne n'est réalisable qu'avec l'algorithme itératif, et le théorème garantit son efficacité quelque soit le graphe [6].

Ce théorème est donc aussi élégant qu'utile. L'idée de la preuve est étonnamment simple : on montre que la loi de transition (5) définit une application $T: p \mapsto p'$ qui est contractante de rapport $1 - c$. Le résultat découle ainsi du théorème du point fixe de Banach [7].

Conclusion

Pour être utile, un moteur de recherche doit non seulement énumérer les résultats d'une requête mais les classer par ordre d'importance. Or, estimer la pertinence des pages web est un profond défi de modélisation.

En première approximation Google analyse le graphe formé par les liens entre pages web. Interprétant un lien $j \rightarrow i$ comme « vote » de la page P_j en faveur de la page P_i , le modèle PageRank (6) définit une mesure de « popularité ».

Le théorème du point fixe assure que cette équation admet une unique solution, et justifie l'algorithme itératif (5) pour l'approcher. Celui-ci est facile à implémenter sur ordinateur et assez efficace pour les graphes de grande nature.

Muni de ces outils mathématiques et d'une habile stratégie d'entreprise, Google gagne des milliards de dollars. Il fallait y penser !

Annexe — quelques pistes d'approfondissement



Le modèle PageRank est-il plausible ?

La structure caractéristique des documents hypertextes sont les citations mutuelles : l'auteur d'une page web ajoute des liens vers les pages qu'il considère utiles ou intéressantes.

L'hypothèse à la base du modèle PageRank est que l'on peut interpréter un lien comme un vote ou une recommandation. Des millions d'auteurs de pages web 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 sera à débattre et à analyser plus en détail. L'ultime argument en faveur du modèle PageRank, par contre, est son succès : le classement des résultats semble bien refléter les attentes des utilisateurs.

Le classement 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. Après avoir compris l'algorithme de Google, les concepteurs de sites web appliquent cette connaissance afin d'améliorer leur classement... Dans un premier temps, il suffit d'attirer des liens, de préférence ceux émis des pages importantes, et il vaut mieux en émettre très peu, de manière bien choisie.

Ces stratégies et astuces sont devenues un domaine très actif, dit « search engine optimization » (SEO). Cette évolution rend l'évaluation des pages web encore plus difficile : comme l'approche et l'importance de Google sont mondialement connues, les liens s'utilisent différemment de nos jours.

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.

Reformulation matricielle du comptage récursif

L'équation (3) n'est rien d'autre qu'un système d'équations linéaires. Plus explicitement, pour tout couple d'indices $i, j \in \{1, \dots, n\}$, on définit a_{ij} par $a_{ij} := \frac{1}{L_j}$ si $j \rightarrow i$, et par $a_{ij} := 0$ sinon.

On obtient ainsi une matrice $A = (a_{ij})$, et notre équation d'équilibre (3) s'écrit comme

$$m = Am \tag{7}$$

ou encore

$$(A - I)m = 0, \tag{8}$$

ce qui est un honnête système linéaire à n équations et n inconnues m_1, \dots, m_n . (Ici I est la matrice d'identité.)

Dans notre exemple miniature discuté ci-dessus, A est la matrice 12×12 suivante :

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

Comme nous l'avons énoncé ci-dessus, dans cet exemple l'équation $m = Am$ admet comme solution le vecteur (colonne)

$$m = (2, 1, 1, 1, 3, 1, 2, 1, 2, 1, 1, 1)^{\dagger}.$$

Promenade aléatoire sur un graphe

Bien que nous n'utilisons que des arguments d'algèbre linéaire et un peu d'analyse dans \mathbb{R}^n , nous ne nous priverons pas du vocabulaire stochastique, car c'est le point de vue et le langage naturel de notre développement.

Par définition, notre matrice $A = (a_{ij})$ vérifie

$$a_{ij} \geq 0 \quad \text{pour tout } i, j \text{ et } \sum_i a_{ij} = 1 \quad \text{pour tout } j,$$

ce que l'on appelle une matrice stochastique. (La somme de chaque colonne vaut 1, mais on ne peut en général rien dire sur la somme dans une ligne.) Nous supposons ici que toute page émet des liens. Ce n'est pas une restriction sérieuse : si jamais une page n'émet aucun lien on peut la faire pointer vers elle-même.

Nous interprétons a_{ij} comme la probabilité d'aller de la page P_j à la page P_i , en suivant un des ℓ_j liens au hasard. La marche aléatoire associée consiste à se balader sur le graphe suivant les probabilités a_{ij} .

Supposons qu'un vecteur $x \in \mathbb{R}^n$ vérifie

$$x_j \geq 0 \quad \text{pour tout } j \text{ et } \sum_j x_j = 1,$$

ce que l'on appelle un vecteur stochastique ou une mesure de probabilité sur les pages P_1, \dots, P_n : on interprète x_j comme la probabilité de se trouver sur la page P_j .

Effectuons un pas dans la marche aléatoire : avec probabilité x_j on démarre sur la page P_j , puis on suit le lien $j \rightarrow i$ avec probabilité a_{ij} . Ceci nous fait tomber sur la page P_i avec une probabilité $a_{ij}x_j$. Au total, la probabilité d'arriver sur la page P_i par n'importe quel lien est la somme

$$y_i = \sum_j a_{ij}x_j.$$

Autrement dit, un pas dans la marche aléatoire correspond à l'application linéaire

$$T: \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad x \mapsto y = Ax.$$

La marche aléatoire partant d'une probabilité initiale x^0 est l'itération de la transition $x^{t+1} = T(x^t)$ pour $t \in \mathbb{N}$.

Une mesure de probabilité m vérifiant $m = T(m)$ est appelée une mesure invariante ou une mesure stationnaire ou encore une mesure d'équilibre. En termes d'algèbre linéaire (7) c'est un vecteur propre associé à la valeur propre 1. En termes d'analyse, c'est un point fixe de l'application T . C'est ce dernier point de vue que nous privilégions ici.

Le modèle PageRank sous forme matricielle

Dans le modèle PageRank la loi de transition (5) se formalise comme l'application affine

$$T: \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad x \mapsto c\epsilon + (1-c)Ax. \quad (9)$$

Ici le vecteur stochastique $\epsilon = (\frac{1}{n}, \dots, \frac{1}{n})$ correspond à l'équiprobabilité sur toutes les pages, et A est la matrice stochastique définie par le graphe. La constante $c \in [0, 1]$ est un paramètre du modèle.

Restreinte aux vecteurs stochastiques, l'application T est donnée par

$$T(x) = cEx + (1-c)Ax$$

où E est la matrice dont tous les coefficients valent $\frac{1}{n}$. Effectivement, sur le sous-espace affine des vecteurs $x \in \mathbb{R}^n$ vérifiant $\sum_j x_j = 1$ nous avons $Ex = \epsilon$. La restriction de T coïncide donc avec l'application induite par la matrice stochastique $A_c = cE + (1-c)A$. Ainsi nous pouvons appliquer directement la théorie des matrices stochastiques au modèle PageRank.

Chaînes de Markov et ergodicité

Ce que nous venons d'étudier sont des chaînes de Markov, à temps discret et ici à espace d'états fini. En plus nos chaînes de Markov sont homogènes dans le sens que la loi de transition ne change pas au cours du temps.

Le choix du paramètre $c \in]0, 1]$, qui gère la téléportation sur le graphe, garantit que notre chaîne de Markov est irréductible et apériodique. Dans cette situation on a toujours convergence vers une unique mesure stationnaire m : les puissances A^t , où $t \in \mathbb{N}$, convergent vers la matrice dont chaque colonne est m . En particulier, la mesure $x^t = A^t x^0$ converge vers m indépendamment de la mesure initiale x^0 .

Dans cette situation dite « ergodique » la loi des grands nombres est en vigueur : la moyenne « en temps » d'une observable h le long d'une trajectoire est égale à sa moyenne « en espace ». Plus précisément, pour presque toute trajectoire $(\omega_t)_{t \in \mathbb{N}}$ on a l'égalité

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T h(\omega_t) = \sum_j h(j) m_j.$$

En particulier, si h est la fonction caractéristique de la page P_i , alors m_i est la fréquentation moyenne de la page P_i durant la marche aléatoire. Ceci justifie notre interprétation que les pages avec une grande probabilité m_i sont les plus fréquentées, autrement dit les plus populaires.



Grandes matrices creuses

Une implémentation sérieuse de l'algorithme itératif nécessite quelques préparations. Rappelons que la matrice A représentant le graphe du web est très grande, à savoir de l'ordre de quelques milliards de lignes et de colonnes. Comment est-ce possible ?

La manière usuelle de stocker une matrice de taille $n \times n$ est un grand tableau de n^2 coefficients indexés par $(i, j) \in \{1, \dots, n\}^2$. Il est envisageable de stocker ainsi une matrice 1000×1000 , c'est-à-dire un million de coefficients mais ceci est hors de question pour une matrice $n \times n$ où $n \approx 10^6$, voire $n \approx 10^9$. L'approche naïve est donc prohibitive pour le modèle PageRank.

Dans notre cas la plupart des coefficients de la matrice valent zéro car une page n'émet que quelques douzaines de liens typiquement. Dans ce cas, il suffit de stocker les coefficients non nuls, dont le nombre est d'ordre n et non n^2 . Une telle matrice est appelée creuse (ou sparse en anglais).

Pour des applications réalistes, il est donc nécessaire d'implémenter des structures de données et des méthodes adaptées aux matrices creuses. La méthode du point fixe est faite sur mesure pour ce genre d'application, et la loi de transition (5) est facile à implémenter, voir l'article **Comment fonctionne Google ?** cité en bas de page.

Notes

[▲1] Citons d'abord l'article des deux fondateurs de Google, Sergey Brin et Lawrence Page : **The Anatomy of a Large-Scale Hypertextual Web Search Engine** (document pdf de 20 pages), Stanford University 1998. Une présentation mathématique se trouve dans l'article de Kurt Bryan et Tanya Leise : **The \$825,000,000,000 eigenvector : the linear algebra behind Google** (document pdf de 11 pages), SIAM Review 48 (2006) 569—581, ainsi que dans l'article de Rebecca Wills : **Google's PageRank : The Math Behind the Search Engine** (document pdf de 15 pages), Mathematical Intelligencer 28 (2006) 6—11.

[▲2] J'avoue que la première figure présente une certaine ambiguïté. Pour éviter toute confusion je précise que nous regardons ici uniquement le graphe « hypertexte » formé des pages html et de leurs liens mutuels — et non le graphe « physique » formé des ordinateurs et de leurs connexions par câbles. Ce dernier est sans doute intéressant sous d'autres aspects mais ne joue aucun rôle dans la suite.

[▲3] Après une première réflexion l'équation (3) semble circulaire : afin de calculer l'importance m_i de la page P_i (à gauche de l'équation) il faudrait d'abord connaître toutes les valeurs m_j des pages P_j qui la citent (à droite de l'équation). Pour calculer celles-ci il faudrait connaître les valeurs des pages qui les citent et ainsi de suite... cela semble inextricable. On verra plus loin comment s'en sortir. Quand on le voit la première fois, c'est un petit miracle !

[▲4] Le théorème énoncé ici pour les graphes reste vrai pour toute matrice stochastique à coefficients positifs : il s'agit d'une version du célèbre **théorème de Perron-Frobenius**.

[▲5] On ignore les chiffres exacts car depuis des années l'entreprise Google se montre plutôt secrète sur tous les détails techniques. Pour citer **l'autoportrait de Google** : « PageRank permet de mesurer objectivement l'importance des pages Web. Ce classement est effectué grâce à la résolution d'une équation de plus de 500

millions de variables et de plus de 2 milliards de termes. Au lieu de compter les liens directs, PageRank interprète chaque lien de la Page A vers la Page B comme un vote par la Page A pour la Page B. PageRank évalue ensuite l'importance des pages en fonction du nombre de votes qu'elles reçoivent. »

[▲6] L'entreprise Google ne précise rien à ce sujet, mais des observations laissent spéculer que la mise à jour du PageRank s'effectue environ une fois par mois. Il est plausible que, même avec la méthode itérative décrite ci-dessus et une programmation hautement optimisée, le calcul du PageRank des milliards de pages web nécessite au moins quelques jours. En tout cas, les valeurs PageRank sont nécessairement précalculées, tout comme l'annuaire des mots-clés, afin de pouvoir répondre à toute requête dans une fraction de seconde.

[▲7] Pour des compléments et des versions étendues de cet article (en pdf) voir la rubrique *Comment fonctionne Google ?* sur la page de Michael Eisermann à l'Institut Fourier, Université de Grenoble. Vous y trouverez en particulier une preuve, de niveau licence, du théorème du point fixe énoncé ci-dessus et une brève discussion des aspects algorithmiques. Une version abrégée est parue dans le magazine *Quadrature* en avril 2008.

► Crédits images

Pour citer cet article : **Michael Eisermann**, **Comment Google classe les pages web**. *Images des Mathématiques*, CNRS, 2009. En ligne, URL : <http://images.math.cnrs.fr/Comment-Google-classe-les-pages.html>