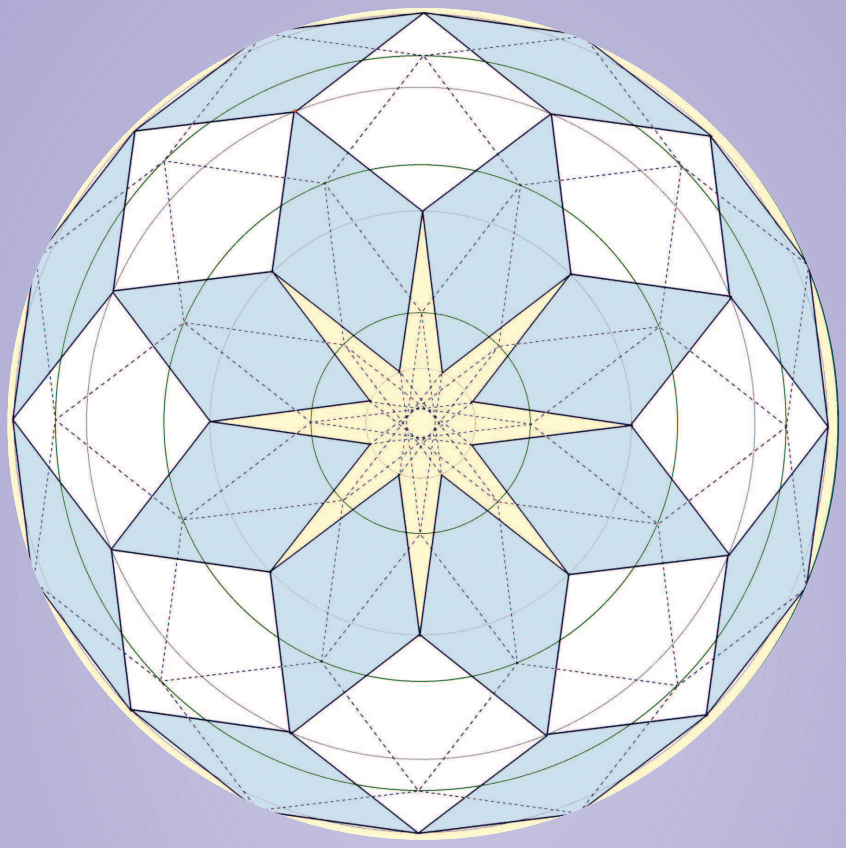
Quadrature

Magazine de mathématiques pures et épicées

La mathématique ouvre plus d'une fenêtre sur plus d'un monde



n° 68
Magazine trimestriel
Avril–Juin 2008
ISSN 1142-2785 – 8 Euros

EDP SCIENCES

Comment fonctionne Google ?

par Michael Eisermann*

Le point fort du moteur de recherche Google est qu'il trie intelligemment ses résultats par ordre d'importance. Nous expliquons ici l'algorithme PageRank qui est à la base de ce classement. Il faut d'abord établir un modèle qui permet de définir ce que l'on entend par « importance ». Une fois ce modèle formalisé, il s'agit de résoudre astucieusement un immense système d'équations linéaires.
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.

Cet article discute les mathématiques utilisées par Google, un moteur de recherche généraliste qui a eu un succès fulgurant depuis sa création en 1998. Le point fort de Google est qu'il trie *par ordre d'importance* les résultats d'une requête, c'est-à-dire les pages web associées aux mots-clés cherchés. L'étonnante efficacité de cette méthode a fait le succès de Google et la fortune de ses fondateurs, Sergey Brin et Lawrence Page. L'idée est née lors de leur thèse de doctorat, puis publiée dans leur article [1]. Il s'agit essentiellement de résoudre un grand système d'équations linéaires et, fort heureusement, l'algorithme itératif qui en découle est aussi simple que puissant. On s'intéresse ici de plus près à cet algorithme, à la fois simple et ingénieux. En conjonction avec une habile stratégie d'entreprise, on pourrait dire que Google gagne des milliards de dollars avec l'algèbre linéaire !

Ajoutons que Google a eu la chance de naître dans une situation favorable, quand la « nouvelle économie » était encore en pleine croissance : le volume d'internet explosait et les moteurs de recherche de première génération avaient du mal à s'adapter aux exigences grandissantes. Si vous voulez en savoir plus sur la foudroyante histoire de l'entreprise Google, ses légendes et anecdotes, vous lirez avec profit le récent livre de David Vise et Mark Malseed [2].

* Institut Fourier, Université Grenoble I, France
URL : www-fourier.ujf-grenoble.fr/~eiserm
e-mail : Michael.Eisermann@ujf-grenoble.fr

I Que fait un moteur de recherche ?

I.1 Fouille de données

À première vue, le principe d'un moteur de recherche est simple : on copie d'abord les pages web concernées en mémoire locale, puis on trie le contenu (les mots-clés) par ordre alphabétique afin d'effectuer des recherches lexiques. Une *requête* est la donnée d'un ou plusieurs mots-clés ; la *réponse* est une liste des pages contenant les mots-clés recherchés. C'est en gros ce que faisaient les moteurs de recherche, dits de première génération, dans les années 1990. Après réflexion, cette démarche simpliste n'est pas si évidente car la quantité des documents à gérer est énorme et rien que le stockage et la gestion efficaces posent des défis considérables. Et cela d'autant plus que les requêtes doivent être traitées en temps réel : on ne veut pas la réponse dans une semaine, mais *tout de suite*.

Une implémentation opérationnelle à cette échelle doit donc employer la force brute d'un réseau puissant, afin de répartir les données et les tâches sur plusieurs ordinateurs travaillant en parallèle. Plus important encore sont les algorithmes, hautement spécialisés et optimisés, sans lesquels même un réseau de quelques milliers d'ordinateurs resterait impuissant devant cette tâche herculéenne.

I.2 Classement des résultats

L'énorme quantité des données entraîne un deuxième problème, plus délicat encore : les pages trouvées sont souvent trop nombreuses, il faut donc

en choisir les plus pertinentes. La grande innovation apportée par Google en 1998 est le tri des pages par ordre d'importance. Ce qui est frappant est que cet ordre correspond assez précisément aux attentes des utilisateurs.

Selon les informations fournies par l'entreprise elle-même, l'index de Google porte sur plus de 8 milliards de documents web. Une bonne partie des informations répertoriées, pages web et documents annexes, changent fréquemment. Il est donc hors de question de les classer manuellement, par des êtres humains : ce serait trop coûteux, trop lent et jamais à jour. L'importance d'une page doit donc être déterminée de manière automatisée, par un algorithme. Comment est-ce possible ?

II Comment mesurer l'importance d'une page web ?

II.1 Le web est un graphe

La particularité des documents *hypertexte* est qu'ils fournissent des liens, des références mutuelles pointant de l'une vers l'autre. Ainsi, on peut considérer le web comme un immense *graphe*, dont chaque page web j est un *sommet* et chaque lien $j \to i$ est une *arête*.

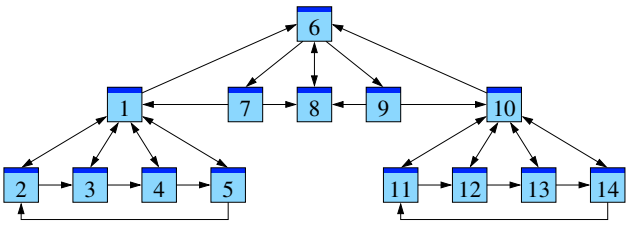


Figure 1. Le web vu comme un graphe.

Dans la suite, on numérote les pages par $1, 2, 3, \ldots, n$ et on écrit $j \to i$ si la page j pointe vers la page i (au moins une fois ; on ne compte pas les liens multiples). Ainsi, chaque page j émet un certains nombre ℓ_j de liens vers des pages « voisines ». À noter que les arêtes sont orientées : si l'on a $j \to i$, on n'a pas forcément le sens inverse $i \to j$. Le graphe de la figure 1, par exemple, s'écrit comme suit :

$$1 \to 2,3,4,5,6\,; \quad 2 \to 1,3\,; \quad 3 \to 1,4\,; \quad 4 \to 1,5\,;$$
$$5 \to 1,3\,; \quad 6 \to 7,8,9\,; \quad 7 \to 8,1\,; \quad 8 \to 6\,;$$
$$9 \to 8,10\,; \quad 10 \to 6,11,12,13,14\,; \quad 11 \to 10,12\,;$$
$$12 \to 10,13\,; \quad 13 \to 10,14\,; \quad 14 \to 10,11.$$

II.2 Comment repérer des pages importantes ?

Dans une première approximation, nous allons négliger le contenu des pages et ne tenir compte que de la structure du graphe.

- Regardons d'abord le groupe des pages $1, 2, 3, 4, 5$. Le dessin suggère que la page 1 sert de racine tandis que les pages $2, 3, 4, 5$ sont subordonnées. Dans ce sens, la page 1 sera sans doute un bon point de départ si vous cherchez des informations.
- Il en est de même pour le groupe $10, 11, 12, 13, 14$, où la page 10 sert de racine alors que $11, 12, 13, 14$ sont subordonnées. À titre d'exemple, il pourrait s'agir d'une page d'accueil et quatre pages annexes, ou d'une introduction et quatre chapitres d'un ouvrage.
- La structure du groupe $6, 7, 8, 9$ est similaire. À noter toutefois que les pages 1 et 10, déjà reconnues comme importantes, font toutes deux référence à la page 6. On pourrait ainsi soupçonner que la page 6 contient de l'information essentielle pour tout l'ensemble.

Heuristiquement, on conclut que les pages $1, 6, 10$ semblent les plus importantes, avec une légère préférence pour la page 6. Soulignons toutefois que notre dessin dans le plan suggère une organisation hiérarchique qui n'est qu'artificielle. Un ordinateur qui analyse cette situation n'a que l'information brute des liens $1 \to 2,3,4,5,6\,;\ 2 \to 1,3\,;$ etc.

Question 1. Est-il possible, par un algorithme, d'associer à chaque page $i = 1, \ldots, n$ une *mesure d'importance* ? Plus explicitement, on souhaite que cette mesure soit un nombre réel $\mu_i \geq 0$ avec la convention que plus μ_i est grand, plus la page i est « importante ».

Remarque 2. La notion d'importance d'une page est nécessairement vague. Qu'est-ce que l'importance ? Peut-il y avoir une mesure objective ? Si oui, comment la définir ? Cette question semble au cœur de toute la problématique. Si vous avez une nouvelle idée pertinente à ce sujet, implémentez-la et devenez riche ! (Ou bien venez en discuter avec moi.)

Dans la suite, notre but sera modeste : le mieux que l'on puisse espérer est que notre analyse dégage un résultat qui *approche* bien l'importance *ressentie* par les utilisateurs. Pour toute application professionnelle, les résultats numériques seront à tester et à calibrer empiriquement.

II.3 Première idée : comptage des liens

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 l'importance μ_i de la page i comme suit :

$$\mu_i = \sum_{j \to i} 1. \tag{1}$$

Interprétation : La somme (1) veut juste dire que μ_i est égal au nombre de liens $j \to i$ reçus par i. C'est facile à définir et facile à calculer : il suffit de compter.

Exemple : Dans notre exemple, les pages 1 et 10 reçoivent 5 liens chacune, alors que la page 6 n'en reçoit que 3. Ainsi $\mu_1 = \mu_{10} = 5$, mais $\mu_6 = 3$ seulement.

Inconvénient : La mesure μ ainsi définie ne correspond pas à l'importance ressentie par les utilisateurs : elle sous-estime l'importance de la page 6.

Manipulation : On peut artificiellement augmenter l'importance d'une page i en créant des pages « vides de sens » pointant vers i. Cette faiblesse fait du comptage une approche peu fiable.

II.4 Seconde idée : comptage pondéré

Certaines pages j émettent beaucoup de liens : ceux-ci sont donc moins spécifiques et, dans un certain sens, leur *poids* est plus faible. Ainsi on pourrait définir une mesure d'importance plus fine comme suit :

$$\mu_i = \sum_{j \to i} \frac{1}{\ell_j}. \tag{2}$$

Interprétation : Comme avant, la somme (2) compte les liens reçus par la page i, mais maintenant chaque lien $j \to i$ n'est compté qu'avec un poids $\frac{1}{\ell_j}$. Il suffit de sommer.

Exemple : Dans notre exemple, on trouve des sommes $\mu_1 = \mu_{10} = 2,5$ et $\mu_6 = 1,4$.

Inconvénient : La mesure μ ainsi définie ne correspond toujours pas bien à l'importance ressentie par les utilisateurs : elle sous-estime à nouveau l'importance de la page 6.

Manipulation : Comme avant, on peut artificiellement augmenter l'importance d'une page i en créant une foule de pages « vides » pointant vers i. De nouveau, la mesure n'est pas fiable.

II.5 Troisième idée : définition récursive

La dernière idée en date, finalement, est celle utilisée par Google. Le principe : *une page i est importante si beaucoup de pages importantes pointent vers i.* Ainsi, on est amené à définir l'importance μ_i de manière récursive comme suit :

$$\mu_i = \sum_{j \to i} \frac{1}{\ell_j} \mu_j. \tag{3}$$

Interprétation : La somme (3) compte chaque lien reçu par i avec poids $\frac{1}{\ell_j}\mu_j$: ceci tient compte de l'importance μ_j de la page d'origine j, et du nombre ℓ_j des liens qui en sont émis.

Exemple : Dans notre exemple, on trouve, après calcul, les valeurs $\mu_6 = 6$ et $\mu_1 = \mu_{10} = 5$, puis $\mu_8 = 4$. Les autres pages suivent avec un grand écart et n'obtiennent que $\mu_i = 2$.

Plausibilité : Les pages 6, 1, 10, 8 sont effectivement repérées comme les plus importantes. Ceci veut dire que la mesure μ ainsi obtenue correspond assez bien à l'importance ressentie par les utilisateurs, comme motivée ci-dessus. (On discutera de la raison au paragraphe V.)

Robustesse : Si l'on ajoute des pages « vides de sens », elles recevront l'importance 0 et ne contribueront pas au calcul. Ainsi, la manipulation évidente n'influence plus le résultat.

L'équation (3) est facile à écrire, mais moins évidente à résoudre : naïvement parlant, pour calculer μ_i il faut d'abord connaître les termes de droite, donc les μ_j, ce qui a l'air circulaire… Notre objectif est donc d'expliquer pourquoi une solution existe et comment la calculer de manière efficace.

II.6 Apparaît l'algèbre linéaire…

Après réflexion, 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

$$a_{ij} := \begin{cases} \frac{1}{\ell_j} & \text{si } j \to i, \\ 0 & \text{sinon.} \end{cases} \tag{4}$$

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

$$\mu = A\mu \qquad \text{ou encore} \qquad (A - I)\mu = 0,$$

ce qui est un honnête système linéaire, que l'on peut résoudre par des méthodes adéquates.

Exemple 3. Dans notre exemple, A est la matrice 14×14 explicitée ci-dessous et l'équation $\mu = A\mu$ admet la solution énoncée. (La vérifier !) C'est même la seule à multiplication par un scalaire près.

$$A = \begin{pmatrix}
0 & 1/2 & 1/2 & 1/2 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/3 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 1/2 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/5 & 0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/5 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/5 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/5 & 0 & 0 & 1/2 & 0
\end{pmatrix},$$

$$\mu = \begin{pmatrix}
5 \\ 2 \\ 2 \\ 2 \\ 2 \\ 6 \\ 2 \\ 4 \\ 2 \\ 5 \\ 2 \\ 2 \\ 2 \\ 2
\end{pmatrix}.$$

Définition 4. Soit $A \in \mathbb{R}^{n\times n}$ une matrice et soit $v \in \mathbb{R}^n \setminus \{0\}$ un vecteur non nul. Si $Av = \lambda v$ pour un scalaire $\lambda \in \mathbb{R}$, alors on dit que v est un *vecteur propre* de la matrice A, associé à la *valeur propre* λ.

Pour notre application, nous nous intéressons donc aux vecteurs propres de A associés à $\lambda = 1$. On montrera plus bas qu'un tel vecteur propre existe et que la solution est essentiellement unique (paragraphe IV.2).

III Marche aléatoire sur la toile

III.1 Matrices stochastiques

Les arguments du paragraphe II nous mènent à étudier une certaine matrice A qui code la structure du web. Avant de résoudre l'équation $A\mu = \mu$, on va essayer d'en développer une intuition. L'idée est de réinterpréter μ comme une mesure de « popularité » des pages web.

Chaque page j émet un certain nombre ℓ_j de liens, ce que l'on code par des coefficients a_{ij} suivant l'équation (4) ci-dessus. Par la suite, nous supposons que $\ell_j \geq 1$, ce qui n'est pas une restriction sérieuse : si jamais une page n'émet pas de liens, on peut la faire pointer vers elle-même.

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

$$\begin{aligned} a_{ij} &\geq 0 && \text{pour tout } i, j \text{ et} \\ \sum_i a_{ij} &= 1 && \text{pour tout } j, \end{aligned}$$

ce que l'on appelle une *matrice stochastique*. (À noter que la somme de chaque colonne vaut 1, mais on ne peut en général rien dire sur la somme dans une ligne.)

On peut interpréter a_{ij} comme la probabilité d'aller de la page j à la page 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}. Notre modèle admet ainsi une étonnante interprétation probabiliste : aussi étrange que cela puisse apparaître, on modélise un surfeur aléatoire, qui ne lit jamais rien mais qui clique au hasard !

Soulignons donc à nouveau que ce n'est pas le contenu des pages web qui est pris en compte pour le calcul de « l'importance », mais uniquement la structure du graphe formé par les pages et les liens entre elles. (Ne vous faites pas trop de souci pour l'instant, on discutera plus bas, au paragraphe V, pourquoi ce modèle est tout de même plausible.)

III.2 Convergence vers une mesure invariante

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

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

ce que l'on appelle un *vecteur stochastique* ou une *mesure de probabilité* sur les pages $1, \ldots, n$: on interprète x_j comme la probabilité de se trouver sur la page j.

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

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

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

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

Remarque 5. Si x est un vecteur stochastique, alors son image $y = Ax$ l'est aussi. Effectivement, $y_i \geq 0$ car $y_i = \sum_j a_{ij}x_j$ est une somme de termes positifs ou nuls et, de plus,

$$\begin{aligned} \sum_i y_i &= \sum_i \sum_j a_{ij}x_j = \sum_j \sum_i a_{ij}x_j \\ &= \sum_j \Big(\sum_i a_{ij}\Big)x_j = \sum_j x_j = 1. \end{aligned}$$

Définition 6. Une mesure de probabilité μ vérifiant $\mu = T(\mu)$ est appelée une *mesure invariante* ou une *mesure d'équilibre*. En termes d'algèbre linéaire, c'est un vecteur propre associé à la valeur propre 1. En termes d'analyse, μ est un point fixe de l'application T.

Exemple 7. Itérer la marche aléatoire avec une probabilité initiale u_0 veut dire que l'on considère les mesures de probabilités successives u_1, u_2, u_3, ... définies par $u_{t+1} = Au_t$. Voici un exemple démarrant sur la page 8, c'est-à-dire $u_0 = (0,0,0,0,0,0,0,1,0,0,0,0,0,0)$:

temps	page 1	page 2	page 3	page 4	page 5	page 6	page 7
t=0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
t=1	0.000	0.000	0.000	0.000	0.000	1.000	0.000
t=2	0.000	0.000	0.000	0.000	0.000	0.000	0.333
t=3	0.167	0.000	0.000	0.000	0.000	0.333	0.000
t=4	0.000	0.033	0.033	0.033	0.033	0.400	0.111
t=5	0.122	0.017	0.017	0.017	0.017	0.111	0.133
t=6	0.100	0.033	0.033	0.033	0.033	0.293	0.037
t=7	0.084	0.036	0.036	0.036	0.036	0.210	0.098
t=8	0.122	0.035	0.035	0.035	0.035	0.168	0.070
t=9	0.105	0.042	0.042	0.042	0.042	0.217	0.056
...							
t=28	0.125	0.050	0.050	0.050	0.050	0.151	0.050
t=29	0.125	0.050	0.050	0.050	0.050	0.150	0.050
t=30	0.125	0.050	0.050	0.050	0.050	0.150	0.050

temps	page 8	page 9	page 10	page 11	page 12	page 13	page 14
t=0	1.000	0.000	0.000	0.000	0.000	0.000	0.000
t=1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
t=2	0.333	0.333	0.000	0.000	0.000	0.000	0.000
t=3	0.333	0.000	0.167	0.000	0.000	0.000	0.000
t=4	0.111	0.111	0.000	0.033	0.033	0.033	0.033
t=5	0.244	0.133	0.122	0.017	0.017	0.017	0.017
t=6	0.170	0.037	0.100	0.033	0.033	0.033	0.033
t=7	0.135	0.098	0.084	0.036	0.036	0.036	0.036
t=8	0.168	0.070	0.122	0.035	0.035	0.035	0.035
t=9	0.126	0.056	0.105	0.042	0.042	0.042	0.042
...							
t=28	0.100	0.050	0.125	0.050	0.050	0.050	0.050
t=29	0.100	0.050	0.125	0.050	0.050	0.050	0.050
t=30	0.100	0.050	0.125	0.050	0.050	0.050	0.050

On observe un phénomène de *diffusion*, très plausible après réflexion :

- On commence au temps $t = 0$ sur la page 8 avec une probabilité de 1.000.
- Au temps $t = 1$, on se trouve sur la page 6 avec une probabilité de 1.000, suivant le seul lien $8 \to 6$.
- Pour $t = 2$, on tombe sur une des pages voisines suivant $6 \to 7, 8, 9$, chacune avec une probabilité de $\frac{1}{3}$.
- Dans les itérations suivantes, la probabilité se propage sur tout le graphe.
  On constate qu'à partir de $t = 5$, la distribution est partout non nulle.
- Après 30 itérations, on est très proche (à 10^{-3} près) de la solution μ déjà exhibée ci-dessus.

On conclut, au moins empiriquement, que la probabilité tend vers notre distribution d'équilibre μ. À noter qu'il ne s'agit pas de l'équiprobabilité : certaines pages sont plus fréquentées que d'autres ! Comme motivé plus haut, ceci reflète bien le rôle particulier des pages 6, 1, 10, 8.

Remarque 8. L'interprétation de la limite $u_t \to \mu$ est la suivante : μ_i est la probabilité de se trouver sur la page i après une très longue marche aléatoire. Ainsi les pages avec une grande probabilité μ_i 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 μ comme indicateur.

III.3 Le modèle PageRank utilisé par Google

Il se trouve que notre modèle a encore un grave défaut quant aux propriétés mathématiques, ainsi qu'à son utilité pratique :

Exemple 9. Le graphe suivant est une variante de l'exemple donné au paragraphe II.1, où s'ajoute la page 15 qui n'émet pas de liens. Pourtant, le résultat diffère drastiquement : la seule mesure invariante est $\mu = (0, \ldots, 0, 1)$, car notre surfeur aléatoire tombera tôt ou tard sur la page 15, où il demeure pour le reste de sa vie. Ce résultat ne reflète évidemment pas l'importance des pages, qui devrait rester inchangée (ou presque).

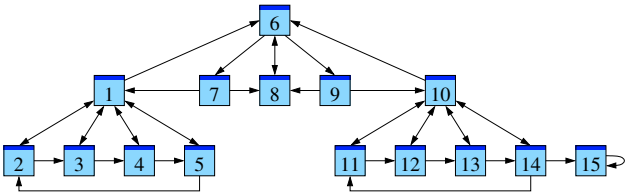


Figure 2. Une variante du graphe initial.

Pour cette raison, Google utilise un modèle plus raffiné, dépendant d'un paramètre $c \in [0,1]$:

- Avec probabilité c, le surfeur abandonne la page actuelle et recommence sur une des n pages du web, choisie de manière équiprobable.
- Avec probabilité $1 - c$, le surfeur suit un des liens de la page actuelle j, choisi de manière équiprobable parmi tous les ℓ_j liens émis. (C'est la marche aléatoire discutée ci-dessus.)

Cette astuce évite en particulier de se faire piéger par une page sans issue. Plus généralement, elle garantit d'arriver n'importe où dans le graphe, indépendamment des questions de connexité.

Ce nouveau modèle probabiliste se formalise comme l'application affine

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

Ici, A est la matrice stochastique définie par l'équation (4). Le vecteur stochastique $\varepsilon = (\frac{1}{n}, \ldots, \frac{1}{n})$ correspond à l'équiprobabilité sur toutes les pages. La constante $c \in [0,1]$ est un paramètre du modèle.

Remarque 10. La valeur $\frac{1}{c}$ est le *nombre moyen* de liens suivis avant de recommencer sur une page aléatoire. En général, on choisira la constante c positive mais proche de zéro. Par exemple, $c = 0.15$ correspond à suivre 7 liens en moyenne. (On pourrait argumenter que ceci correspond empiriquement au comportement des utilisateurs… À débattre.)

IV Existence et unicité d'une solution

IV.1 Le théorème du point fixe

Une fonction $f\colon \mathbb{R} \to \mathbb{R}$ est *contractante* de rapport $k < 1$ si elle vérifie $|f(x) - f(y)| \leq k|x - y|$ pour tout $x, y \in \mathbb{R}$. Sous cette hypothèse, f admet un et un seul point fixe $\mu \in \mathbb{R}$, $f(\mu) = \mu$, et pour tout $u_0 \in \mathbb{R}$, la suite itérative $u_{m+1} = f(u_m)$ converge vers μ.

C'est exactement l'argument qu'il nous faut pour notre application. On a déjà vu, d'ailleurs, que la convergence se produisait dans notre exemple ci-dessus. Est-ce une coïncidence ? Non, c'est encore une manifestation du fameux théorème du point fixe. Comme nous travaillons sur les vecteurs $x \in \mathbb{R}^n$, nous sommes amenés à le généraliser convenablement :

Définition 11. Pour un vecteur $x \in \mathbb{R}^n$, on définit sa *norme* par $|x| := \sum_i |x_i|$. (C'est une honnête norme qui a toutes les bonnes propriétés usuelles.) Une fonction $f\colon \mathbb{R}^n \to \mathbb{R}^n$ est dite *contractante* de rapport $k < 1$ (par rapport à la norme $|\cdot|$) si elle vérifie $|f(x) - f(y)| \leq k|x - y|$ pour tout $x, y \in \mathbb{R}^n$.

Théorème 12 (le théorème du point fixe). *Si $f\colon \mathbb{R}^n \to \mathbb{R}^n$ est une fonction contractante de rapport $k < 1$, alors :*

1. *Il existe un et un seul point $\mu \in \mathbb{R}^n$ vérifiant $f(\mu) = \mu$.*
2. *Pour toute valeur initiale $u_0 \in \mathbb{R}^n$, la suite itérative $u_{m+1} = f(u_m)$ converge vers μ.*
3. *On a $|u_m - \mu| \leq k^m|u_0 - \mu|$, la convergence vers μ est donc au moins aussi rapide que celle de la suite géométrique k^m vers 0. Pour le calcul concret, on a l'estimation de l'écart*

$$|u_m - \mu| \leq \frac{k}{1-k}|u_m - u_{m-1}|.$$

Remarque 13. Dans la pratique, on ignore souvent la limite μ, mais on peut facilement calculer la suite itérative u_m. Pour contrôler la qualité de l'approximation u_m, on majore l'écart $|u_m - \mu|$ entre u_m et la limite inconnue par la quantité $\frac{k}{1-k}|u_m - u_{m-1}|$, très facile à calculer.

Démonstration. Comme il s'agit d'une très belle preuve, je ne peux m'empêcher de la refaire ici.

Unicité. — Si $x, y \in \mathbb{R}^n$ sont deux points fixes d'une fonction f qui est contractante de rapport $k < 1$, alors $|x - y| = |f(x) - f(y)| \leq k|x - y|$. Ceci n'est possible que pour $|x - y| = 0$, donc $x = y$.

Existence. — Une récurrence facile montre que $|u_{m+1} - u_m| \leq k^m|u_1 - u_0|$ pour tout $m \in \mathbb{N}$, puis

$$\begin{aligned} |u_{m+p} - u_m| &\leq |u_{m+p} - u_{m+p-1}| + \ldots + |u_{m+1} - u_m| \\ &\leq (k^{p-1} + \ldots + k^0)|u_{m+1} - u_m| \\ &= \tfrac{1-k^p}{1-k}|u_{m+1} - u_m| \end{aligned}$$

pour tout $m, p \in \mathbb{N}$. La suite (u_m) est donc de Cauchy et converge puisque $(\mathbb{R}^n, |\cdot|)$ est complet. Notons $\mu := \lim u_m$ sa limite et vérifions qu'il s'agit d'un point fixe. Puisque f est contractante, elle est continue. L'équation de récurrence $u_{m+1} = f(u_m)$ donne donc

$$\mu = \lim u_{m+1} = \lim f(u_m) = f(\lim u_m) = f(\mu).$$

Vitesse de convergence. — Pour tout u_0 la suite itérative $u_{m+1} = f(u_m)$ vérifie $|u_m - \mu| \leq k^m|u_0 - \mu|$, donc $u_m \to \mu$. On a déjà établi la majoration $|u_{m+p} - u_m| \leq \frac{k}{1-k}|u_m - u_{m-1}|$, et le passage à la limite $p \to \infty$ donne l'inégalité cherchée. □

IV.2 Application au modèle PageRank

Proposition 14. *Soit $A \in \mathbb{R}^{n\times n}$ une matrice stochastique et $T\colon \mathbb{R}^n \to \mathbb{R}^n$ l'application définie par*

$$T(x) = c\varepsilon + (1 - c)Ax$$

avec une constante $c \in]0, 1]$. Alors l'application T est contractante de rapport $k = 1-c < 1$. Par conséquent, elle admet une unique mesure invariante $\mu = T(\mu)$ et, pour tout vecteur initial u_0, la suite itérative $u_{m+1} = T(u_m)$ converge vers le point fixe μ, avec la vitesse énoncée ci-dessus.

Démonstration. Il suffit de prouver que T est contractante, de rapport $k = 1 - c < 1$, pour faire appel au théorème du point fixe. Regardons deux vecteurs $x, y \in \mathbb{R}^n$ et essayons de majorer $z := Tx - Ty$ en fonction de $|x - y|$. On a $z = kA(x - y)$ donc

$z_i = k \sum_j a_{ij}(x_j - y_j)$ pour tout $i = 1, \dots, n$. Ceci nous permet de calculer la norme :

$$\begin{aligned} |Tx - Ty| = |z| &= \sum_i |z_i| = \sum_i \Big| k \sum_j a_{ij}(x_j - y_j) \Big| \\ &\leq k \sum_i \sum_j |a_{ij}(x_j - y_j)| \\ &= k \sum_j \sum_i a_{ij} |x_j - y_j| \\ &= k \sum_j \Big(\sum_i a_{ij} \Big) |x_j - y_j| = k|x - y|. \end{aligned}$$

Ceci prouve que $T : \mathbb{R}^n \to \mathbb{R}^n$ est contractante de rapport k comme énoncé. L'application T admet donc un unique point fixe $\mu \in \mathbb{R}^n$. Remarquons finalement que le point fixe est un vecteur stochastique, c'est-à-dire qu'il satisfait $\mu_i \geq 0$ et $\sum_i \mu_i = 1$: si l'on démarre avec un vecteur stochastique u_0, alors tous les itérés u_m restent stochastiques, donc leur limite μ l'est aussi. (Exercice.) □

Remarque 14. La proposition inclut le cas trivial $c = 1$: dans ce cas $T(x) = \varepsilon$ est constante, donc $x = \varepsilon$ est l'unique point fixe. Dans l'autre extrême, on pourrait considérer $c = 0$, mais $T = A$ n'est pas forcément contractante. Par exemple, pour un graphe à n sommets sans arêtes entre eux, nous obtenons la matrice identité $A = I$, qui admet tout vecteur $x \in \mathbb{R}^n$ comme point fixe. Un bon choix de c se situe donc quelque part entre 0 et 1 (voir la remarque 10).

Remarque 15. Le fait que la solution soit unique est fondamental : une fois que le modèle est établi, le théorème nous garantit une unique mesure μ, sans équivoque. Mieux encore, la suite itérative converge toujours vers μ, indépendamment du point de départ. En l'absence de toute autre information, on pourra donc démarrer avec $u_0 = \varepsilon = (\frac{1}{n}, \dots, \frac{1}{n})$ pour calculer la limite $u_m \to \mu$.

Remarquons à ce propos que Google est obligé de mettre à jour ses données régulièrement, car le web change sans cesse. Disons que Google met à jour le vecteur μ chaque semaine. Pour ce calcul, il serait maladroit de recommencer par $u_0 = \varepsilon$! Il est sans doute plus avantageux de recycler l'information déjà obtenue : on choisira $u_0 = \mu_{\text{ancien}}$, la mesure de la semaine d'avant. Ainsi, peu d'itérations suffiront pour réajuster μ, en supposant que le graphe n'est que légèrement modifié.

Remarque 16. La proposition précédente se généralise au théorème de Perron-Frobenius : si une matrice réelle A a tous ses coefficients positifs, $a_{ij} > 0$ pour $i, j = 1, \dots, n$, alors le rayon spectral de A est donné par une valeur propre $\lambda \in \mathbb{R}_+$ et l'espace propre associé E_λ est de dimension 1. De plus, la matrice A admet un vecteur propre $v \in E_\lambda$, dont tous les coefficients sont positifs.

L'algorithme itératif correspondant est souvent appelé la « méthode de la puissance ». Il se généralise à une matrice A quelconque et permet d'approcher numériquement un vecteur propre v associé à la valeur propre λ de module $|\lambda|$ maximal, pourvu que cette valeur propre soit unique et simple.

V Le modèle est-il plausible ?

La structure caractéristique d'un document *hypertexte* sont les *liens* vers d'autres documents. L'auteur d'une page web ajoute ainsi des liens vers les pages qu'il considère utiles ou « importantes ». Autrement dit, on peut interpréter un lien comme un *vote* ou une *recommandation*. Or, il ne suffit pas de compter les liens, car ils n'ont pas tous le même poids. Nous avons donc raffiné notre heuristique : une page est importante si beaucoup de pages importantes pointent vers elle. Cette définition peut sembler circulaire, mais le développement mathématique ci-dessus montre comment s'en sortir par le théorème du point fixe.

V.1 Hypothèses implicites

Des millions d'auteurs de pages web lisent et jugent mutuellement leurs pages, puis leurs jugements s'expriment par les liens qu'ils mettent sur leurs pages. Le modèle de la marche aléatoire en profite en transformant l'évaluation mutuelle en une mesure globale de popularité. D'après l'autoportrait de Google, « la technologie de Google utilise l'intelligence collective du web pour déterminer l'importance d'une page. » Nous venons de voir comment cette phrase peut s'interpréter mathématiquement. Une triple hypothèse y est implicite :

(1) Les liens reflètent fidèlement les appréciations des *auteurs* des pages web.

(2) Ces appréciations correspondent bien à celles des *lecteurs* des pages web.

(3) Le modèle du surfeur aléatoire les traduit fidèlement en une mesure de popularité.

En soutien de ces hypothèses, on mentionne parfois la « nature démocratique » du web pour dire que les lecteurs et les auteurs ne font qu'un et que l'échange des informations est libre. C'est une idéalisation de moins en moins plausible, surtout quant à l'aspect commercial du web. En 1993, seul 1,5 % des sites web étaient dans le domaine `.com` ; en 2003, ils représentaient

plus de 50 % du web et la fréquentation des pages devrait donner des proportions similaires.

V.2 Descriptif ou normatif ?

Le statut de Google lui-même a complètement changé :

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

En réalité il est devenu normatif : Aujourd'hui, son écrasant succès fait de Google une référence normative : si une page figure en tête du classement, alors elle est importante.

À titre d'illustration, citons un exemple devenu classique. Le mathématicien français Gaston Julia, né le 3 février 1893, devint célèbre pour ses contributions à la théorie de fractales, largement popularisée par son élève Benoît Mandelbrot depuis les années 1970. Pour son anniversaire le 3 février 2004, la page d'accueil de Google montrait une variation fantaisiste du logo usuel. Un clique dessus lançait la recherche d'images associées aux mots-clés « Julia » et « fractale ». Deux des pages en tête du classement étaient hébergées à un institut de l'université de Swinburne, à Melbourne en Australie. Comme tous les jours, des millions d'internautes ont visité la page de Google et, ce jour-là, une certaine fraction a suivi le lien du logo, pour tomber sur la page à Swinburne. Ce trafic soudain a suffi pour submerger le serveur australien, qui rendit l'âme aussitôt. Les images fractales durent être déplacées et une page explicative fut mise à la place [3]. Elle conclut par une question mémorable (d'après Job **1** 21) :

> Google giveth, and Google taketh away, blessed be Google ?
> [Google avait donné, Google a repris, que le nom de Google soit béni ?]

V.3 Peut-on manipuler Google ?

Pour des sites web commerciaux, l'optimisation de leur classement est devenue un enjeu important. Évidemment, le fournisseur d'un service commercial souhaite que son site soit le plus visité possible et ceci passe par Google : des millions de clients potentiels utilisent Google et suivent typiquement les liens en tête du classement. Comment améliorer son classement, son importance calculée par Google ? Voici ce qu'en dit l'entreprise Google :

> Les méthodes complexes et automatiques utilisées par les recherches Google rendent quasi impossible toute manipulation humaine des résultats. (…) Google ne pratique pas la vente des positions dans ces résultats ; autrement dit, il n'est pas possible d'acheter une valeur PageRank supérieure à la réalité du Web.

Pourtant, afin d'améliorer son classement par Google, il suffit d'attirer des liens, de préférence ceux émis par des pages importantes, et il vaut mieux en émettre très peu, de manière bien choisie. Les stratégies et astuces sont elles-mêmes devenues un domaine très actif, dit « *search engine optimization* » (SEO). Ceci confirme, en particulier, que l'omniprésence de Google change l'utilisation des liens par les auteurs… ce qui remet en question l'hypothèse à la base même du modèle.

V.4 Comment évolue Google ?

L'algorithme de base que nous venons de décrire fut mis en œuvre en 1998 et reste, semble-t-il, le fondement de l'efficacité légendaire de Google. (D'ailleurs, la méthode a été brevetée par l'université de Stanford en 2001, ainsi qu'une version raffinée en 2004, et le nom « PageRank » est une marque déposée de Google Inc. [4].)

Évidemment, le modèle du surfeur aléatoire n'est qu'une première approximation. La méthode actuellement utilisée a sans doute été adaptée et peaufinée au fil des années, afin de rendre le classement encore plus utile, c'est-à-dire plus proche des attentes des utilisateurs, et plus robuste contre des tentatives de manipulations. Contrairement à l'algorithme de base, toutes les modifications ultérieures restent un secret de l'entreprise Google.

Toujours est-il que les webmestres les plus inventifs arrivent souvent à influencer le classement en leur faveur pour se positionner sur les premières pages des résultats. En réaction, Google est obligé d'améliorer son algorithme pour rattraper les tricheurs, au moins les plus flagrants. C'est l'habituelle course du gendarme et du voleur, mais typiquement Google ne s'en sort pas trop mal.

Effectivement, Google a tout intérêt à maintenir la bonne qualité de ses résultats afin de défendre sa popularité qui, rappelons-le, est la source de ses revenus. Si l'on veut y voir un aspect positif, on pourrait dire que cette éternelle compétition fait évoluer les moteurs de recherche.

Remerciements

Je tiens à remercier mon collègue Tanguy Rivoal pour ses conseils linguistiques et surtout son encouragement à publier ces notes de cours sous forme d'article de vulgarisation.

Références

[1] S. Brin et L. Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Stanford University, 1998. (Pour trouver ce texte en ligne cherchez-le avec Google.)

[2] D. Vise et M. Malseed, *Google story*, Dunod, Paris, 2006.

[3] *The power of Google*, `local.wasp.uwa.edu.au/~pbourke/fractals/quatjulia/google.html`.

[4] Wikipedia, *PageRank*, `en.wikipedia.org/ wiki/PageRank` `fr.wikipedia.org/ wiki/PageRank`. (La page francophone attend toujours une rédaction digne du sujet.)