Par Michael Eisermann

L'algorithme PageRank de Google
Promenade sur la toile

Depuis plus d'une décennie, Google domine le marché des moteurs de recherche sur Internet.
Son point fort ? Il trie intelligemment ses résultats par ordre de pertinence. Son fonctionnement repose en fait sur une judicieuse modélisation mathématique.

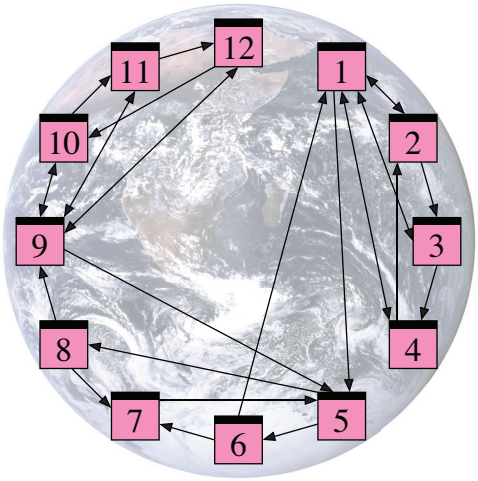


Illustration de la structure de graphe du Web.

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 et est disponible en ligne. Le pilier du succès de Google est une judicieuse modélisation mathématique.

+ 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. Toute tentative de classification semble vouée à l'échec, d'autant plus que le Web évolue rapidement : une multitude d'auteurs ajoutent constamment de nouvelles pages et modifient 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 en mémoire locale et trie les mots par ordre alphabétique. Le résultat est un annuaire de mots clés avec leurs pages Web associées.

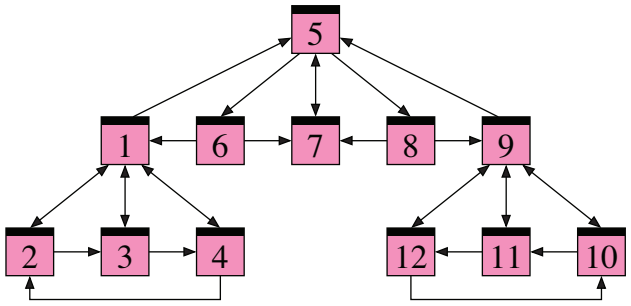
Pour un mot clé donné, il y a typiquement des milliers de pages correspondantes (plus d'un million pour « tangente » par exemple). Comment aider l'utilisateur à repérer les résultats potentiellement intéressants ? C'est sur ce point que Google a apporté sa grande innovation.

+ La Toile 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, dont la figure suivante montre un exemple en miniature.
Dans toute la suite, notons les pages Web par P_1, P_2… P_n et écrivons $j \to i$ si la page P_j cite la page P_i. Dans notre graphe, nous avons un lien $1 \to 5$, par exemple, mais pas de lien $5 \to 1$.

+ Pages pertinentes

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 $j \to i$ 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 et de la pertinence de la page P_i.
Analysons notre exemple sous cet aspect. La présentation suivante de notre graphe suggère une hiérarchie possible – encore à justifier.



Le graphe des liens entre les différentes pages.

Parmi les pages P_1, P_2, P_3 et 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} et P_{12} où la page P_9 sert de référence commune. La structure du groupe P_5, P_6, P_7 et 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 donc la plus pertinente.

+ 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 l'importance m_i de la page P_i comme le nombre des liens $j \to i$. En formule, ceci s'écrit comme suit :

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

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 $m_5 = m_7 = 3$.
Mais ce qui est pire encore, c'est que ce comptage naïf est trop facile à manipuler en ajoutant des pages sans intérêt qui recommandent une page quelconque.

+ 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 :

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

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 $m_5 = 3/2$ et $m_7 = 4/3$. Et, comme dans la situation précédente, ce comptage est très facile à truquer.

+ Comptage récursif

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

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

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

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

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 le comptage récursif 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.

Un peu d'aléa

Avant de tenter de résoudre le système linéaire, 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 1/3. Voici les probabilités suivantes (à 0,001 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}
$t=0$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$t=1$	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$t=2$	0,000	0,000	0,000	0,000	0,000	0,333	0,333	0,333	0,000	0,000	0,000	0,000
$t=3$	0,167	0,000	0,000	0,000	0,333	0,000	0,333	0,000	0,167	0,000	0,000	0,000
$t=4$	0,000	0,042	0,042	0,042	0,417	0,111	0,111	0,111	0,000	0,042	0,042	0,042
$t=5$	0,118	0,021	0,021	0,021	0,111	0,139	0,250	0,139	0,118	0,021	0,021	0,021
...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$t=29$	0,117	0,059	0,059	0,059	0,177	0,059	0,117	0,059	0,117	0,059	0,059	0,059
$t=30$	0,117	0,059	0,059	0,059	0,177	0,059	0,117	0,059	0,117	0,059	0,059	0,059

On observe une diffusion qui converge assez rapidement vers une distribution stationnaire. 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}
$t=0$	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$t=1$	0,000	0,250	0,250	0,250	0,250	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$t=2$	0,375	0,125	0,125	0,125	0,000	0,083	0,083	0,083	0,000	0,000	0,000	0,000
$t=3$	0,229	0,156	0,156	0,156	0,177	0,000	0,083	0,000	0,042	0,000	0,000	0,000
$t=4$	0,234	0,135	0,135	0,135	0,151	0,059	0,059	0,059	0,000	0,010	0,010	0,010
$t=5$	0,233	0,126	0,126	0,126	0,118	0,050	0,109	0,050	0,045	0,005	0,005	0,005
...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$t=69$	0,117	0,059	0,059	0,059	0,177	0,059	0,117	0,059	0,117	0,059	0,059	0,059
$t=70$	0,117	0,059	0,059	0,059	0,177	0,059	0,117	0,059	0,117	0,059	0,059	0,059

Bien que la diffusion mette plus de temps, 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 pour lesquelles m_i est le plus grand sont les plus « fréquentées » ou les plus « populaires ». Dans la quête de classer les pages Web, c'est encore un argument pour utiliser la mesure m comme indicateur.

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 $j \to i$ est alors p_j/l_j. La probabilité d'arriver au temps $t + 1$ sur la page P_i est donc :

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

Étant donnée la distribution initiale p, cette loi de transition 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 définissant m par un comptage récursif.

Trous noirs

Que se passe-t-il quand notre graphe contient une page (ou un groupe de pages) sans issue ? Pour illustration, reportez-vous au graphique de la page suivante.
L'interprétation comme marche aléatoire permet de résoudre l'équation définissant m par un comptage récursif sans aucun calcul. La page P_{13} absorbe toute la probabilité car notre surfeur aléatoire tombera tôt ou tard sur cette page, où il demeurera pour le reste de sa vie. Ainsi la solution est $m = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$. Notre modèle n'est donc pas encore satisfaisant.

Téléportation

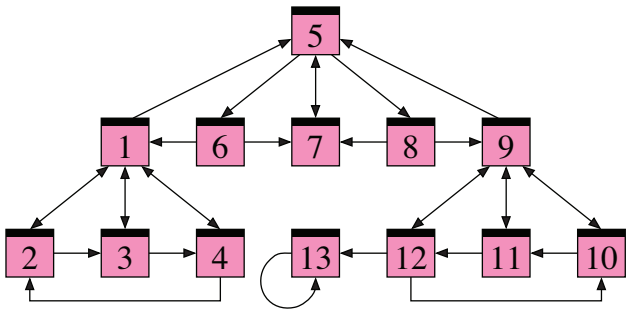
Pour échapper aux trous noirs, Google utilise un modèle plus raffiné : avec une probabilité fixée c, le surfeur abandonne sa page actuelle P_j et recommence sur l'une des n pages du Web, choisie de manière équiprobable. Sinon, avec probabilité $1 - c$, le surfeur suit l'un des liens de la page P_j, choisi de manière équiprobable. 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 :

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

Le premier terme c/n provient de la téléportation, le second terme est la marche aléatoire précédente. La mesure d'équilibre vérifie donc :

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

Le paramètre c est encore à calibrer. Pour $c = 0$, nous obtenons le modèle précédent. Pour c dans l'intervalle]0 ; 1], la valeur $1/c$ est le nombre



Modification du graphe.

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). C'est ce modèle qui est appelé PageRank. Par exemple, le choix $c = 0{,}15$ correspond à suivre environ six liens en moyenne, ce qui semble une description réaliste. Pour conclure l'analyse de notre exemple, le tableau de la marche aléatoire partant de la page P_1peut être visualisé ci-contre.:
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 $m_1 = m_9 = 0{,}12$.

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

Avec quelques outils mathématiques et une habile stratégie d'entreprise, Google gagne des milliards.

+ PageRank

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 :
Considérons un graphe fini quelconque et fixons le paramètre c dans $]0\ ;\ 1]$. Alors l'équation vérifiée par la mesure d'équilibre m admet une unique solution vérifiant $m_1 + \dots m_n = 1$. Dans cette solution, $m_1, \dots m_n$ sont tous positifs. Pour toute distribution de probabilité initiale, le processus de diffusion p' 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.
La démonstration de ce résultat découle d'une application du théorème du point fixe.

Pour être utile, un moteur de recherche doit non seulement énumérer les résultats d'une requête, mais également 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 $j \to i$ comme « vote » de la page P_j en faveur de la page 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'utilisation de p' (méthode itérative) pour l'approcher. Cet algorithme itératif est facile à implémenter et assez efficace pour les graphes de grandeur nature. Muni de ces outils mathématiques et d'une habile stratégie d'entreprise, Google gagne des milliards de dollars. Il fallait y penser !

M.E.

RÉFÉRENCE

- **Jacques Bair**, *Comment Google classe les pages ?*, *Tangente Sup* n° 44–45, pp. 10–13, septembre–octobre 2008.
- **Michael Eisermann**, *Comment fonctionne Google ?*, www-fourier.ujf-grenoble.fr/~eiserm/enseignement\\#google", 15 pages, 2008.