



Mathématiques et moteur de recherche

Michael Eisermann

Institut für Geometrie und Topologie, Universität Stuttgart, Allemagne

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 : le pilier de son succès est une judicieuse modélisation mathématique.

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. 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 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 de cinq millions pour *matrice*, par exemple). 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.

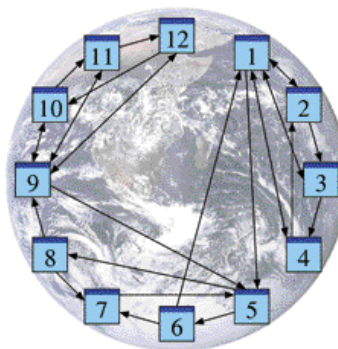
Dans la suite je note les pages web par $P_1, P_2, P_3, \dots, P_n$ et j'écris $j \rightarrow i$ si la page

P_j site la page P_i . Dans notre graphe nous avons un lien $1 \rightarrow 5$, par exemple, mais pas de lien $5 \rightarrow 1$.

Comment exploiter ce graphe?

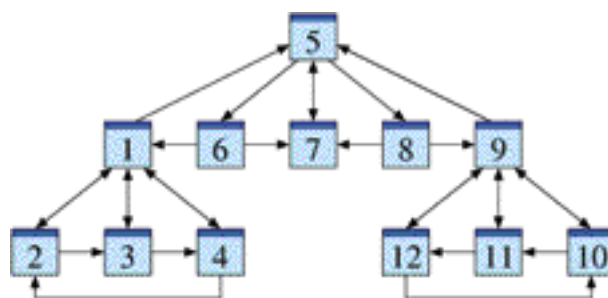
Les liens sur internet ne sont pas aléatoires mais ont été édités avec soin. Quels renseignements pourrait nous donner ce graphe ?

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. A 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

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 P_i comme le nombre des liens $j \rightarrow i$. En formule ceci s'écrit comme suit:

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

Autrement dit, μ_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 $\mu_1 = \mu_9 = 4$ devant $\mu_5 = \mu_7 = 3$. 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:

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

Autrement dit, μ_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 $\mu_1 = \mu_9 = 2$ devant $\mu_5 = 3/2$ et $\mu_7 = 4/3$. 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 μ_i de manière récursive comme suit:

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

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

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

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

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 ?

A titre d'exemple, supposons que notre surfeur démarre au temps $t = 0$ sur la page P_7 . Le seul lien pointe vers P_7 , 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:

	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. 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, la mesure stationnaire est la même! Elle coïncide d'ailleurs avec notre solution $\mu = (2,1,1,1,3,1,2,1,2,1,1,1)$, ici divisée par 17 pour normaliser la somme à 1. Les pages où μ_i est 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 μ comme indicateur.

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 pro-

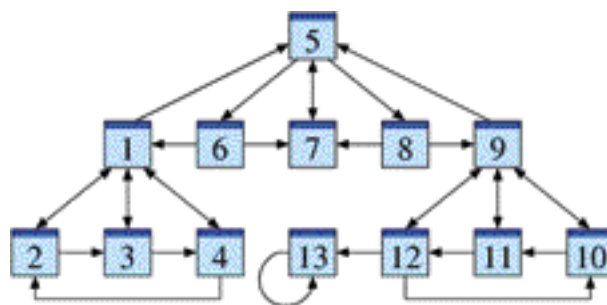
babilité de partir de P_j et de suivre le lien $j \rightarrow i$ est alors P_{ji}/ℓ_j . La probabilité d'arriver au temps $t + 1$ sur la page P_i est donc

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

Etant 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 $\mu = T(\mu)$, qui est justement notre équation (3)

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 modifié :



L'interprétation comme marche aléatoire permet de résoudre l'équation (3) sans aucun calcul : la page P_3 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 $\mu = (0,0,0,0,0,0,0,0,0,0,0,0,1)$. Notre modèle n'est donc pas encore satisfaisant.

Le modèle 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.

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

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

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

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

Le paramètre c est encore à calibrer. Pour $c = 0$ nous obtenons le modèle précédent. Pour $0 < c \leq 1$ la valeur $1/c$ 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).

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_i :

	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 $\mu_5 = 0.15$ avant les pages P_7 et P_9 avec $\mu_7 = \mu_9 = 0.12$.

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 :

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 $\mu_1 + \dots + \mu_n = 1$. Dans cette solution $\mu_1, \dots, \mu_n$ sont tous positifs. Pour toute distribution de probabilité initiale le processus de diffusion (5) converge vers cette unique mesure stationnaire μ . La convergence est au moins aussi rapide que celle de la suite géométrique $(1 - c)^n$ vers 0.*

L'idée de la preuve est simple : on montre que la loi de transition (5) définit une application $T : p \rightarrow p'$ qui est contractante de rapport $1 - c$. Le résultat découle ainsi du théorème du point fixe de Banach.

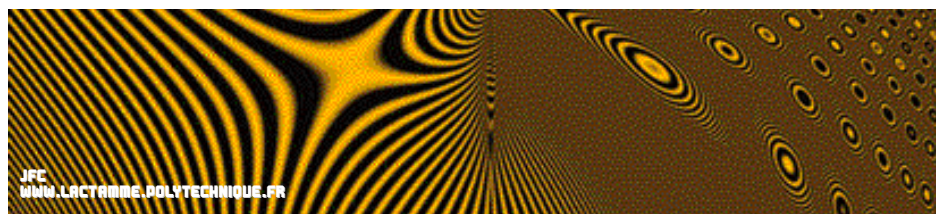
Conclusion

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 Page-Rank (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 et assez efficace pour les graphes de grande taille.

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



Pour en savoir (un peu) plus

S. Brin, L. Page : *The Anatomy of a Large-Scale Hypertextual Web Search Engine.*
Stanford University 1998, (20 pages)
<http://infolab.stanford.edu/pub/papers/google.pdf>

M. Eismann : *Comment Google classe les pages web.* *Images des Mathématiques*, CNRS, 2009.
<http://images.math.cnrs.fr/Comment-Google-classe-les-pages.html>