Autour de l'année 1995, une bonne dizaine de moteurs de recherche s'offraient aux internautes. Mais le nombre de sites de la grande toile explosait et, de plus en plus, l'efficacité de ces moteurs diminuait. En 1998 Google lançait son moteur et, en quelques années, tous les internautes l'avaient adopté. Pourquoi ?

Pourquoi utilisez-vous Goog

Yvan Saint-Aubin
Université de Montréal

Les technologies ignorées

Un enfant découvre les objets qui l'entourent avec plaisir et curiosité. En vieillissant il utilise de nombreux outils, sans réaliser les prouesses technologiques que, souvent, ces objets représentent. Nos mères n'ont jamais eu à utiliser les planches à laver grâce aux machines à laver, la génération des trente ans a joui des disques compacts sans les égratignures des vinyles, les aventuriers modernes ont délaissé les cartes topographiques plastifiées pour leur gps et ceux nés après 1990 ne peuvent comprendre pourquoi certaines personnes ont encore des lignes téléphoniques par câble à leur domicile.

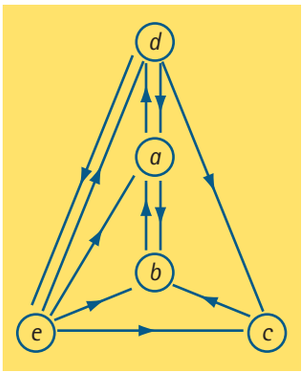
Je dois donc ouvrir ce texte par une explication du « miracle » qu'est le moteur de recherche offert par Google. Les moteurs de recherche sont un peu l'index d'une encyclopédie. La seule différence est que, contrairement aux divers articles d'une encyclopédie, les sites de la grande toile respectent peu de règles. Ces sites peuvent contenir du texte, une photo, un clip sonore, une vidéo ou même un mélange de ces divers éléments. Leur qualité n'est soumise à aucun contrôle : elle peut être excellente (visitez par exemple le site officiel du Prix Nobel) ou nulle ! Leur rythme de progression a été exponentiel jusqu'à récemment et il semble inutile d'obtenir un consensus au sein de la communauté des internautes sur ce qui est important. Les premiers moteurs de recherche fonctionnaient comme l'index d'une bonne encyclopédie, ignorant les caractéristiques propres à la grande toile. Suite à une requête sur le nom propre « Gauss », ces moteurs vous présentaient pêle-mêle des centaines de pages : la Plomberie Gauss, le régime amaigrissant du Dr Gauss, le centre commercial Gauss Plaza, etc. Bien audacieux était l'internaute qui prétendait y trouver une biographie du mathématicien Carl Friedrich Gauss ! (Faites une recherche maintenant avec le mot « gauss » !) Avec la croissance de la grande toile, ces moteurs traditionnels devenaient de plus en plus inutiles. De nouveaux outils étaient nécessaires et les pionniers de la navigation internet se mirent à la recherche d'une « boussole » pour la grande toile.

Cette « boussole » apparaît sous forme d'un algorithme simple, non pas pour limiter le nombre de sites présentés suite à une requête, mais bien pour les ordonner en mettant en premier les pages plus utiles. Mais comment trouver les pages « utiles » si la communauté ne peut en venir à un consensus sur la qualité des sites ? C'est que l'algorithme dû à L. Page, S. Brin, R. Motwani et T. Winograd sonde la communauté... d'une façon détournée ! (Les deux premiers auteurs quitteront l'Université Stanford, où ils avaient entrepris un doctorat, pour fonder la compagnie Google.)

le?

Le promeneur impartial

Pour décrire l'algorithme original du moteur de recherche de Google, j'utiliserai une métaphore, celle du promeneur impartial. Pour comprendre sa promenade une toute petite toile est bien suffisante, par exemple celle ci-contre qui ne contient que 5 sites étiquetés *a*, *b*, *c*, *d* et *e*.



On peut s'amuser à imaginer que *a* est le site officiel des Jeux Olympiques, *b* celui des Jeux d'hiver 2014 de Sochi, *c* celui de l'agence antidopage, *d* l'office de tourisme russe et *e* un grand périodique. Cette dernière page désire informer ses lecteurs; elle contient des liens vers le site des jeux de Sochi, mais aussi vers les pages *b*, *c* et *d*; ces liens sont indiqués par les flèches partant de *e*, comme $e \to a$, $e \to b$, etc. Le site *b* des Jeux Olympiques est plus sobre et ne contient qu'un lien vers le site de Sochi (*a*) indiqué par la flèche $b \to a$. Bref, sur cette microtoile, les lettres représentent les sites et une flèche entre elles, de la forme $i \to j$, indique que le site *i* contient un lien amenant au site *j*.

Un promeneur est posé en une page de la toile et se voit donner la directive de changer de site à chaque minute, tout en préservant son « impartialité ». Cette dernière consigne le force donc à choisir équitablement entre les liens qui s'offrent à lui. S'il est en *c*, il ne peut qu'emprunter le lien menant à *b* et c'est donc en *b* qu'il se retrouvera au pas suivant. Si cependant son pas le plus récent l'a mené à la page *e*, quatre pages s'offrent maintenant à lui et il devra donc choisir « impartialement », par exemple par tirage au sort. Il se retrouvera donc au pas suivant en *a*, *b*, *c* ou *d*, avec probabilité 1/4 pour chacune de ces destinations. À cause de ces choix qu'il doit faire, le chemin suivi par le promeneur impartial n'est plus déterministe. Il est quand même possible de décrire sa trajectoire, mais de façon probabiliste.

Supposons qu'au départ, c'est-à-dire au pas 0, le promeneur est à la page *d*. Nous dirons (de façon un peu pédante) qu'il est en *d* avec probabilité 1. Si p_i^t dénote la probabilité de trouver le promeneur au site *i* (un des *a*, *b*, *c*, *d*, *e*) à son pas *t*, cette position initiale se résume donc aux égalités :

$$p_d^0 = 1 \text{ et } p_a^0 = p_b^0 = p_c^0 = p_e^0 = 0.$$

Trois liens s'offrent à lui pour son prochain pas, des liens qui le mènent vers les pages *a*, *c* et *e*. Puisqu'il est impartial, il choisira équitablement entre les trois et la probabilité de le trouver en un de ces trois sites est :

$$p_a^1 = p_c^1 = p_e^1 = 1/3 \text{ et } p_b^1 = p_d^1 = 0.$$

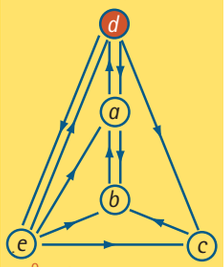
Nous avons ajouté les probabilités de le trouver en *b* ou *d*, probabilités qui sont nulles, puisqu'aucun chemin ne mène de *d* à *b* ou à *d*.

Le calcul des probabilités au second pas est un peu plus compliqué. Étudions d'abord les promenades qui se terminent en *a*. Le promeneur est avec probabilité non nulle en *a*, *c* ou *e*, après le premier pas. De ces trois sites il ne peut atteindre *a* qu'en provenance de *e*. Or, de *e*, le promeneur choisira parmi quatre sites, chacun avec probabilité 1/4 par impartialité, et le chemin allant de *d* au pas 0, à *e* au pas 1, et à *a* au pas 2 a donc probabilité

$$p_a^2 = 1/3 \cdot 1/4 = 1/12.$$

Il vaut la peine de répéter l'exercice pour les promenades qui se terminent en *b*. À nouveau, au pas 1, le promeneur peut être en *a*, *c* ou *e*. Il peut atteindre le site *b* à partir de ces trois sites.

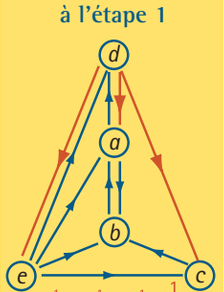
Probabilités à l'étape 0



$p_d^0 = 1$

$p_a^0 = p_b^0 = p_c^0 = p_e^0 = 0$

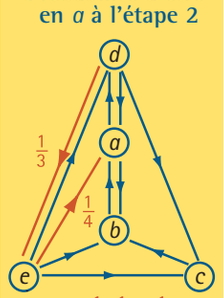
Probabilités à l'étape 1



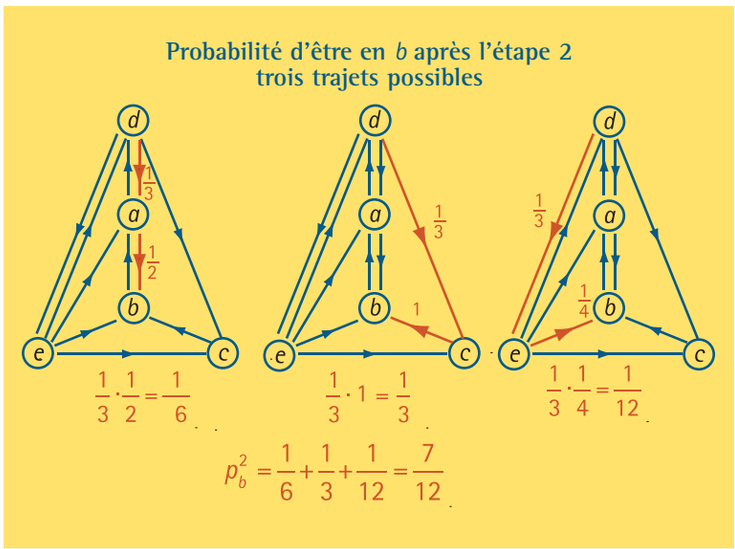
$p_a^1 = p_c^1 = p_e^1 = \frac{1}{3}$

$p_b^1 = p_d^1 = 0$

Probabilité d'être en *a* à l'étape 2



$p_a^2 = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$



Par conséquent, trois trajectoires contribuent maintenant :

$d \to a \to b$ (probabilité = $1/3 \cdot 1/2$),

$d \to c \to b$ (probabilité = $1/3 \cdot 1$) et

$d \to e \to b$ (probabilité = $1/3 \cdot 1/4$).

La probabilité de le trouver en b au pas 2 est donc

$$p_b^2 = 1/6 + 1/3 + 1/12 = 7/12.$$

Les autres probabilités pour le second pas sont calculées similairement :

$$p_a^2 = 1/12,\ p_b^2 = 7/12,\ p_c^2 = 1/12,$$
$$p_d^2 = 1/4,\ p_e^2 = 0.$$

Cette méthode peut être répétée pour obtenir les probabilités au troisième pas :

$$p_a^3 = 2/3,\ p_b^3 = 1/8,\ p_c^3 = 1/12,$$
$$p_d^3 = 1/24,\ p_e^3 = 1/24$$

ou à un pas ultérieur. Mais cet exercice perd son attrait rapidement. Remarquons quand même que la somme des cinq probabilités au pas 1 (ou au pas 2 ou 3) est égale à 1, comme il se doit : le promeneur est sûrement en un des cinq sites ! (Ceux qui connaissent le produit matriciel voudront lire l'encadré qui explique une façon simple de calculer les probabilités à un pas t quelconque.)

Des questions probabilistes

Plutôt que de calculer les probabilités pour les pas ultérieurs, essayons plutôt de répondre à quelques questions de nature probabiliste.

- Que deviennent les probabilités p_a^t, p_b^t, p_c^t, p_d^t, p_e^t après plusieurs pas ? Par exemple, est-ce que p_a^t devient à peu près constant lorsque t est très grand, c'est-à-dire est-ce que p_a^{100} et $p_a^{1\,000}$ sont à peu près les mêmes nombres ?
- Est-ce que ces cinq probabilités changent beaucoup si le promeneur démarre son périple à partir d'un autre site, par exemple le site c plutôt que d ?
- Où est-il le plus probable de trouver le promeneur en $t = 1\,000$ et en $t = 1\,000\,000$?

Même si cette dernière question semble ardue, il est possible de tenter une réponse. Avant de vous laisser deviner, voici quelques observations. Remarquez d'abord que les sites a et c reçoivent, tous les deux, des flèches des sites d et e. Cependant, alors que c ne reçoit que ces deux flèches, le site a en reçoit également une de b. Il est raisonnable de prédire que $p_a^t > p_c^t$ pour de grands t. Une autre observation est que, si le promeneur est en c au pas t, il sera en b au pas suivant. Puisque b reçoit des visites d'autres sites, il est également raisonnable de prédire que $p_b^t > p_c^t$ pour de grands t. Pouvez-vous deviner en quel site le promeneur sera avec la plus grande probabilité après de nombreux pas, en admettant qu'un tel site existe ? (Le prochain paragraphe révèle la réponse. Arrêtez-vous donc quelques minutes pour y réfléchir...)

Des réponses probabilistes

Les probabilités de trouver le promeneur en un des cinq sites se stabilisent rapidement. Après cent pas, ces probabilités sont, à quatre chiffres après la virgule décimale:

$$p_a^{100} = 0{,}3666,\ p_b^{100} = 0{,}2834,$$
$$p_c^{100} = 0{,}0833,\ p_d^{100} = 0{,}2000$$
$$\text{et } p_e^{100} = 0{,}0666.$$

En fait, il est possible de montrer que, quel que soit le site de départ de la promenade, ces probabilités tendront après de nombreux pas vers les cinq nombres

$$\pi_a = 11/30,\ \pi_b = 17/60,\ \pi_c = 1/12,$$
$$\pi_d = 1/5 \text{ et } \pi_e = 1/15$$

et que l'endroit le plus probable où trouver le promeneur sera donc le site *a*, suivi de près par le site *b*. Ces probabilités « ultimes » sont appelées les *probabilités asymptotiques*. Ces probabilités ont une belle propriété. Si on les utilise comme probabilité initiale ou au pas *t*, les probabilités au pas suivant et tout pas ultérieur seront identiques : si $p_a^0 = \pi_a$, alors $p_a^t = \pi_a$ pour tous les pas *t* suivants, et des égalités similaires valent pour les autres sites. Cette propriété remarquable permet de déterminer ces probabilités asymptotiques.

L'idée cruciale (et lumineuse) des concepteurs est que le comportement asymptotique du promeneur indique une préférence des internautes. Par exemple le fait que le site *a* soit visité plus souvent ou avec une plus grande probabilité que le site *c* est révélateur : cela signifie que la majorité des gens qui ont construit les sites de cette microtoile ont jugé plus important d'ajouter un lien vers *a* que, disons, vers *c*. Ainsi, il y a un consensus entre internautes pour dire que, parmi les sites ayant trait aux Jeux Olympiques de Sochi, le site *a* est plus important que le site *c*. Si vous lanciez une requête sur les mots « jeux sochi », alors le microGoogle vous présenterait les cinq pages dans l'ordre décroissant de leur probabilité :

$$\pi_a > \pi_b > \pi_d > \pi_c > \pi_e$$

et les pages apparaîtraient donc dans l'ordre :

a, *b*, *d*, *c*, *e*

puisque la communauté a « voté » pour cet ordre d'importance. C'est de cette façon que Google ordonne, avant de vous les présenter, les pages trouvées suite à votre requête. Morale de cette histoire : créer un lien vers un site, c'est voter pour ce site !

De la microtoile à 5 sites à la grande toile

Que reste-t-il de ces observations si la petite toile du promeneur est remplacée par la grande toile avec ses quelque 10^{13} sites entrelacée par les nombreuses flèches qui les lient? Sous des hypothèses assez générales sur les flèches (les liens entre les sites), l'existence de probabilités asymptotiques caractérisant la trajectoire après de nombreux pas est assurée. L'indépendance du point initial du promeneur et l'unicité du comportement asymptotique sont plus délicates. (Voir l'encadré *Le théorème de Frobenius*.)

Mais en pratique...

Rappelez-vous que la métaphore du promeneur impartial décrit l'algorithme *original* du moteur de recherche de Google. Plusieurs changements y ont été apportés. Cependant, même pour l'algorithme original, les ingénieurs de Google ont dû partager le travail entre plusieurs parcs d'ordinateurs.

Un premier parc d'ordinateurs se consacre à explorer la grande toile, comme le promeneur impartial, enregistrant à chaque pas, le contenu du site visité (les mots qui s'y trouvent) et vers quels autres sites il pointe. Ce premier parc construit donc le graphe comme celui de la microtoile où les sommets sont les sites et les liens les flèches qui les joignent.

Le 25 juillet 2008, Google annonça que leur moteur de recherche avait maintenant indexé 10^{12} sites distincts ! Depuis je ne me préoccupe plus des nouveaux records atteints. Ce qui est sûr, c'est que le calcul du comportement asymptotique du promeneur, ces nombres π_i qui seront utilisés pour ordonner les sites de votre requête, représente un des plus gros problèmes d'algèbre matricielle présentement résolus sur la planète. Un second parc d'ordinateurs est consacré à sa résolution et il

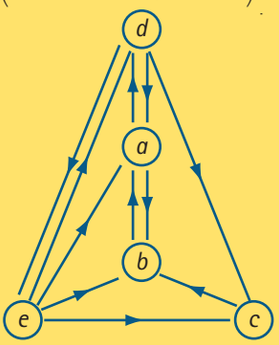
l'accomplit mensuellement à partir des données obtenues par le premier parc d'ordinateurs. Ajoutons cependant que des calculs complémentaires permettent maintenant d'indexer de nouvelles pages rapidement. Par exemple, l'annonce des Prix Nobel en octobre dernier a été faite d'abord par les blogues des grands périodiques. Ces blogues furent répertoriés en quelques minutes.

Finalement, un troisième parc d'ordinateurs attend nos requêtes, trouve les pages qui contiennent nos mots-clés et, surtout, les ordonne selon le rang décrit par les probabilités π_i calculées le mois précédent par le second parc d'ordinateurs.

Les probabilités du promeneur impartial par le calcul matriciel

$$T = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1/3 & 1/4 \\ 1/2 & 0 & 1 & 0 & 1/4 \\ 0 & 0 & 0 & 1/3 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/3 & 0 \end{pmatrix} \end{array}$$

Départ (colonnes) — Arrivée (lignes)



Le fastidieux calcul des probabilités p_i^t de trouver le promeneur au site i au pas t gagne à être formulé matriciellement. La première étape consiste à capturer toute l'information de la petite toile à 5 sites dans une matrice T carrée 5×5. Les sites sont ordonnés alphabétiquement (a, b, c, d, e) et chacun correspond dans cet ordre à une ligne et une colonne de la matrice T. Chacune des colonnes est alors construite comme suit. La première colonne indique les destinations possibles du site a. Chaque site accessible à partir de a est indiqué dans la ligne correspondante de la colonne a par 1/2. (Si trois flèches quittaient a, alors le chiffre utilisé serait 1/3, et ainsi de suite.) Les sites qui ne peuvent être atteints de a sont notés par un zéro. En répétant ce calcul pour chaque site (et donc chaque colonne), la matrice T ci-contre est obtenue. La seconde étape consiste simplement à regrouper les probabilités p_i^t, $i \in \{a; b; c; d; e\}$, en un vecteur colonne p^t à cinq composantes. Alors les probabilités au pas $t+1$ sont liées à celles au pas t par le simple produit matriciel

$$p^{t+1} = Tp^t.$$

Cette équation est très simple. Pour comprendre comment cette élégante relation donne le même résultat que le calcul de probabilités de chacun des chemins fait dans le texte, il est utile de prendre un exemple. Calculons la probabilité p_b^2 que le promeneur soit au site b au second pas. Seuls les sites a, c et e pointent vers le site b. Il faudra donc multiplier les probabilités p_a^1, p_c^1 et p_e^1 par les probabilités que le promeneur quitte un de ces sites pour b. Ces probabilités p_a^1, p_c^1 et p_e^1 sont contenues dans le vecteur colonne $p^{t=1}$. La probabilité p_b^2 vient du produit matriciel de la seconde ligne de T avec le vecteur $p^{t=1}$ qui donne précisément

$$p_b^2 = \frac{1}{2}\cdot p_a^1 + 0\cdot p_b^1 + 1\cdot p_c^1 + 0\cdot p_d^1 + \frac{1}{4}\cdot p_e^1$$
$$= \frac{1}{6} + 0 + \frac{1}{3} + 0 + \frac{1}{12} = \frac{7}{12}.$$

Cette formulation est d'autant plus élégante qu'elle permet de caractériser le comportement asymptotique. Puisque le comportement asymptotique ne change pas d'un pas à l'autre, il faut que le vecteur π qui contient les probabilités π_a, π_b, ... satisfasse $T\pi = \pi$. C'est cette équation que résolvent mensuellement plusieurs ordinateurs à la compagnie Google. Évidemment, la matrice T qu'ils étudient est celle décrivant la grande toile. C'est donc une matrice carrée $n \times n$ avec un $n > 10\,000\,000\,000\,000$.

Et Google aujourd'hui ?

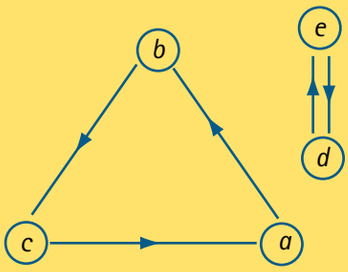
L'algorithme original décrit par le promeneur impartial demeure utilisé (nous croyons), mais de nombreuses améliorations lui ont été apportées. Nous ne les connaissons pas, secret professionnel oblige, mais nous pouvons en constater les effets. L'une est l'indexation rapide de nouvelles pages dont il a été question ci-dessus. Une seconde est l'ordre des mots-clés de la requête qui influence l'ordre de présentation des sites trouvés. Une troisième est l'utilisation de l'origine géographique de la requête qui, à nouveau, change l'ordre des sites trouvés. Ainsi le 8 octobre dernier, *google.ca* m'aiguillait vers les périodiques canadiens pour découvrir les gagnants du Prix Nobel de physique, mais *google.com.au* orientaient ses utilisateurs vers des périodiques australiens.

Malgré toutes ces améliorations, les requêtes continuent à être traitées en moins d'une seconde... une prouesse technologique que nous oublions parfois lorsque nous explorons la grande toile à l'aide du moteur de recherche de Google.

Le théorème de Frobenius

L'existence d'un comportement asymptotique pour le promeneur impartial (et de son unicité) est régie par le théorème de Frobenius, un grand résultat issu de l'algèbre linéaire et de l'analyse. Pour expliquer l'algorithme du promeneur, j'ai choisi une toile qui possède un et un seul comportement asymptotique représentant bien les « votes » de la communauté internaute. Cependant les hypothèses du théorème de Frobenius ne sont pas nécessairement vérifiées pour la grande toile.

Plutôt que décrire les hypothèses techniques de ce théorème, voici deux petites toiles qui soulignent certaines des difficultés. La première toile est en fait composée de deux parties qui ne sont pas interconnectées.



Les sites e *et* d *sont inaccessibles à partir de* a.

Si le promeneur démarre au site *a* de la toile à gauche, il ne visitera jamais ni *d* ni *e*. De plus, sa promenade demeure complètement déterministe et ne s'approche pas d'un des « bons » comportements asymptotiques qui est

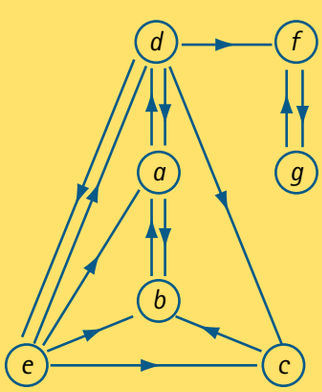
$$\pi_a = \pi_b = \pi_c = 1/3 \quad \text{et} \quad \pi_d = \pi_e = 0\,.$$

La seconde toile, en bas à droite, décrit une difficulté qui se présente sûrement dans la grande toile. À chaque fois que le promeneur visite le site *d* de cette toile, il a une chance sur quatre d'opter pour le site *f*. Tôt ou tard, c'est ce qu'il fera et alors il sera capturé sur le petit ilôt créé par les sites *f* et *g*. À partir de ce moment fatidique, sa promenade sera

$$f \to g \to f \to g \to \ldots$$

et elle ignorera le reste de la toile. Clairement les probabilités asymptotiques de cette promenade caractérisent mal l'importance relative des pages de cette toile.

Les ingénieurs de Google doivent trouver des méthodes pour contrer les difficultés que représentent ces petites toiles et pour obtenir des probabilités π_a, π_b, ... représentant correctement nos « votes ». Une méthode efficace avait été proposée dans l'article présentant l'algorithme original. D'autres ont peut-être été ajoutées.



En allant en f, *le promeneur devient captif de l'ilôt formé par les sites* f *et* g.