Printemps des Sciences 2008

Google, graphes et matrices...

La matrice cachée de Google (Michel Rigo, ULg)

Dossier pédagogique, 5ème – 6ème secondaire

Inutile de le présenter : Google est le moteur de recherche le plus connu et le plus utilisé au monde. Mais comment les concepteurs de Google font-ils pour classer les milliers de pages se rapportant à un mot-clé donné, de façon telle que les pages les plus représentatives occupent toujours les premières positions du classement ? Ce tour de force repose sur de véritables résultats mathématiques combinant théorie des graphes et algèbre linéaire.

Table des matières

1 Comment classer des pages web ? 1

2 Position du problème : le modèle de L. Page et S. Brin 2

3 Et les matrices, ça sert à quoi ? 4

4 Recherche d'une solution et perturbation du modèle 5

5 Interprétation probabiliste 8

6 Et les équipes de Basket aussi ! 11

7 Appendice sur le calcul matriciel 12

Ce dossier pédagogique contient plus de détails que la présentation orale prévue pour le Printemps des Sciences. Bien qu'il soit question de probabilité dans une des sections, ce thème peut sans problème être éludé pour des élèves de 5ième année n'ayant pas encore vu ces notions.

1 Comment classer des pages web ?

La conception d'un moteur de recherche comme Google comporte de nombreux aspects. En effet, il faut tout d'abord parcourir toutes les pages se trouvant sur Internet et les indexer. Ainsi, des "robots" (ou *web-crawler*[1]) sont programmés pour arpenter sans relâche le web et, répertorier et analyser les pages web. Ce texte ne traite nullement de ces sujets (extraction, analyse et gestion d'immenses bases de données). Nous supposerons donc qu'un surfeur, comme vous et moi, a soumis un ou plusieurs mots-clés au moteur de recherche et que ce dernier a déjà extrait de l'ensemble des pages web indexées se trouvant sur Internet, celles faisant référence au sujet voulu. Le problème, bien que simplifié, est encore bien loin d'être résolu comme le montre l'exemple suivant. Lorsqu'on entre le mot clé "`matrice`" ou "`football`", une recherche dans Google affiche :

`Résultats` **`1 - 10`** `sur un total d'environ` **`9 560 000`** `pour` **`matrice`**
`Résultats` **`1 - 10`** `sur un total d'environ` **`279 000 000`** `pour` **`football`**

[1] "to crawl" : ramper, grouiller

Sur plusieurs millions de pages, voire centaines de millions, comment présenter directement les plus intéressantes ? Comment déterminer les pages les plus significatives ? Il faut savoir qu'on estime à près de 10^{10} (10 milliards) le nombre de liens (ou pages) existant sur le web. Sans outils algorithmiques puissants, il ne serait dès lors pas possible d'extraire la moindre information utile.

Le moteur de recherche Google a été développé il y a tout juste dix ans par Sergeï Brin et Larry Page, jeunes doctorants en informatique à l'Université de Stanford. Le nom de Google proviendrait d'une variation du mot '*gogol*' qui signifie (selon certains) 10^{100}. En 1997, le manque d'une véritable méthode efficace de classification se faisait de plus en plus sentir. Pendant la conception de Google au milieu des années 1990, le moteur de recherche le plus populaire de l'époque, Altavista, référençait environ 200 millions de liens et recevait 20 millions de requêtes quotidiennes. Ces chiffres ont depuis été multipliés par plus de dix !

2 Position du problème : le modèle de L. Page et S. Brin

Nous allons présenter, au moyen d'un exemple de taille réduite, le problème qui nous intéresse : *classer des pages pour déterminer les plus importantes ou significatives par rapport à un sujet donné.*

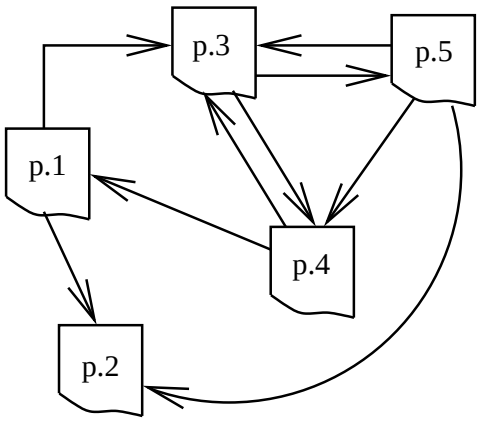

FIG. 1 – Graphe des liens : cinq pages et leurs liens respectifs.

Pour classer ces pages, on leur attribue un score, aussi appelé "*PageRank*"[2]. Plus le score est élevé, plus la page est importante (les pages sont donc classées par "PageRank" décroissant). C'est ce score qui est censé traduire l'importance ou le caractère significatif de la page. Ainsi, l'une des tâches de Google est d'attribuer un "PageRank" à chaque page du web. Ensuite, pour une requête donnée, il "suffit" alors d'extraire, et d'afficher dans l'ordre, les pages en rapport avec la requête.

[2]De manière amusante, on peut traduire "PageRank" par le rang qu'occupe une page, mais on peut aussi y voir la référence à l'un des deux concepteurs de Google.

L'idée à la base du modèle de S. Brin et L. Page tient en deux règles :

R1. On accorde plus d'importance, i.e., un score de "PageRank" plus élevé, aux pages référencées par des pages qui font elles-mêmes autorité dans le domaine, c'est-à-dire qui ont un PageRank élevé ;

R2. On accorde d'autant moins de crédit à une référence, si elle provient d'une page qui dispose de nombreux liens.

Les deux règles sont somme toute assez *naturelles*. Pour la première, au plus une page est référencée par d'autres pages, au plus elle doit faire autorité dans le domaine en question. La deuxième règle sert de "contrepoids" : on ne peut qu'accorder moins de poids, aux sites qui galvaudent, gaspillent, leurs recommandations.

Poser par exemple à mille personnes la question : "*citer les plus grands scientifiques de tous les temps*" (dans cette question, on ne précise pas combien de scientifiques doivent être cités par les participants à l'enquête et chaque personne interrogée peut donner autant de noms qu'elle le désire). Les premiers du classement seront logiquement les scientifiques les plus cités. Cependant accorderez-vous la même importance à la liste fournie par X et qui contient un unique nom : A. Einstein et à la liste fournie par Y qui contient aussi A. Einstein, mais également 499 autres scientifiques. Clairement, X considère A. Einstein comme le plus grand scientifique de tous les temps, par contre, pour Y, il ne s'agit que de l'un des 500 plus grands. Le vote attribué par X doit avoir plus de poids que la voix de Y. Cet exemple permet de mieux sentir l'intérêt de la deuxième règle. Autrement dit, chaque votant possède une unité qu'il peut diviser en autant de parts égales qu'il le désire et il distribue alors ces parts.

La première règle quant à elle voudrait qu'on accorde plus de poids pour la liste fournie par un prix Nobel ou par un scientifique de renom que pour celle d'un citoyen lambda sans qualifications scientifiques particulières.

Reprenons l'exemple de la Figure 1 et appelons s_1, s_2, s_3, s_4, s_5 les scores des pages 1 à 5. Si on ne considère d'abord que la première règle, on peut imaginer, pour rendre compte de **R1**, que le score d'une page s'obtient comme la somme des scores des pages qui pointent vers elle et ainsi obtenir le système suivant :

$$\begin{cases} s_1 = s_4 \\ s_2 = s_1 + s_5 \\ s_3 = s_1 + s_4 + s_5 \\ s_4 = s_3 + s_5 \\ s_5 = s_3 \end{cases} \tag{1}$$

Pour tenir compte de la deuxième règle, on propose de diviser le poids attribué aux liens de chaque page par le nombre de liens de celle-ci (de cette manière, on remplit bien l'objectif d'accorder d'autant moins de poids aux références fournies par une page que celle-ci a de liens). Dans l'exemple du classement des scientifiques donné ci-dessus, cela revient à dire que le vote fourni par X à A. Einstein compte pour 1, alors que le vote de Y compte pour 1/500. On obtient dès lors le système suivant :

$$\begin{cases} s_1 = s_4/2 \\ s_2 = s_1/2 + s_5/3 \\ s_3 = s_1/2 + s_4/2 + s_5/3 \\ s_4 = s_3/2 + s_5/3 \\ s_5 = s_3/2 \end{cases} \tag{2}$$

Nous allons dans les sections suivantes discuter des solutions d'un tel système.

3 Et les matrices, ça sert à quoi ?

En fait, le système (1) peut se réécrire très simplement sous forme matricielle comme

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{pmatrix}}_{\mathbf{s}} = \underbrace{\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{pmatrix}}_{\mathbf{s}}.$$

Autrement dit, rechercher les scores (i.e., la (les) solution(s) du système) revient à trouver un vecteur colonne $\mathbf{s}$ satisfaisant l'équation (matricielle) $\mathbf{As} = \mathbf{s}$.

Remarque. La matrice $\mathbf{A}$ contient toute l'information contenue dans les liens entre les pages donnés à la figure 1. En effet, par exemple, le "1" à la deuxième ligne et première colonne de $\mathbf{A}$ correspond au lien de la page 1 pointant vers la page 2. De même, la deuxième colonne est formée de 0 car la deuxième page ne fournit aucun lien sortant[3]. Cette constatation est tout à fait générale (sans voir la figure, on peut la reconstruire entièrement à partir de la matrice $\mathbf{A}$).

Bien évidemment, c'est plutôt le second système (2) qui nous intéresse puisqu'il modélise les deux règles du modèle de Brin et Page. Pour celui-ci, on obtient une forme matricielle assez proche

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1/3 \\ 1/2 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/3 \end{pmatrix}}_{\mathbf{B}} \underbrace{\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{pmatrix}}_{\mathbf{s}} = \underbrace{\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{pmatrix}}_{\mathbf{s}}.$$

Il suffit en fait de reprendre la matrice $\mathbf{A}$ et de diviser tous les éléments d'une colonne par un même nombre, de façon telle que la somme des éléments de chaque colonne non nulle soit égale à 1. On note $\mathbf{B}$, la matrice ainsi obtenue.

Le problème d'attribuer des scores à chaque page est donc, semble-t-il, bien simple : il suffit de résoudre un système d'équations linéaires comme (2) ou, de manière équivalente, de trouver un vecteur[4] $\mathbf{s}$ tel que

$$\mathbf{Bs} = \mathbf{s}.$$

Il ne faut pas perdre de vue que sur l'exemple traité : 5 équations à 5 inconnues, il n'y a aucune difficulté à résoudre le système. Cependant, pensez qu'avec une situation réelle, des centaines de millions d'équations et d'inconnues peuvent entrer en jeu. Si on désire attribuer un score à l'ensemble des pages de l'Internet, on doit alors résoudre un système de 10^{10} équations et il ne faut pas croire qu'un ordinateur résout instantanément une telle question !

[3] Internet recèle de telles pages ne pointant vers aucune autre, comme par exemple, des images ou des fichiers pdf.

[4] On dit que $\mathbf{s}$ est un *vecteur propre* (de valeur propre 1) de la matrice $\mathbf{B}$.

4 Recherche d'une solution et perturbation du modèle

On peut tout d'abord constater que, si un vecteur colonne $\mathbf{s}$ satisfait l'équation

$$\mathbf{Bs} = \mathbf{s},$$

alors tout multiple de $\mathbf{s}$ la satisfait aussi (on entend par là, que l'on multiplie chaque composante de $\mathbf{s}$ par un même nombre réel). En effet, pour tout nombre réel a, il vient

$$\mathbf{B}(a\mathbf{s}) = a(\mathbf{Bs}) = a\mathbf{s}. \tag{3}$$

Par conséquent, on peut imposer une condition supplémentaire (dite de *normalisation*) demandant que la somme des éléments de $\mathbf{s}$, c'est-à-dire la somme des scores, soit égale à 1. Par exemple, si

$$\mathbf{s} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix},$$

puisque la somme des scores fait 6, il suffit de considérer $\mathbf{s}/6$, qui est encore solution, pour lequel la somme des scores fait 1. Bien évidemment, la solution consistant à prendre tous les scores nuls n'est d'aucun intérêt puisqu'on ne pourrait l'exploiter pour obtenir un classement.

En fait, à ce stade, deux problèmes majeurs se posent. Rien ne garantit qu'un système comme

$$\mathbf{Bs} = \mathbf{s}$$

possède au moins une solution non nulle. Ensuite, si le système possède effectivement une solution (non nulle), il serait intéressant d'en garantir l'unicité (moyennant l'hypothèse de normalisation énoncée plus haut). En effet, si plusieurs solutions sont disponibles, comment donner du sens à la solution qui sera calculée si d'autres solutions, tout aussi valables, mais différentes, peuvent être trouvées. On aurait alors plusieurs classements incomparables... Il s'agit véritablement d'un problème classique maintes fois rencontré en mathématiques, garantir ou prouver l'existence et l'unicité de la solution du problème envisagé.

Si on considère notre exemple, le système (2) ne possède aucune solution non nulle ! (Il ne s'agit que d'un simple exercice un peu plus compliqué que ce que l'on rencontre habituellement dans un cours de 3ième secondaire.) En procédant par "substitution", on trouve

$$\begin{cases} s_1 = s_4/2 \\ s_2 = s_1/2 + s_5/3 \\ s_3 = s_1/2 + s_4/2 + s_5/3 \\ s_4 = s_3/2 + s_5/3 \\ s_5 = s_3/2 \end{cases} \Leftrightarrow \begin{cases} s_1 = s_4/2 \\ s_5 = s_3/2 \\ s_2 = s_4/4 + s_3/6 \\ s_3 = s_4/4 + s_4/2 + s_3/6 \\ s_4 = s_3/2 + s_3/6 \end{cases}$$

La cinquième équation donne $s_4 = 2s_3/3$ et la quatrième $s_4 = 10s_3/9$. De là, on en tire que la seule solution est $s_1 = s_2 = s_3 = s_4 = s_5 = 0$.

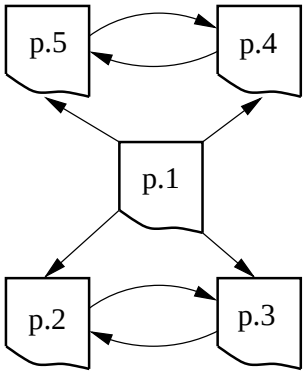

FIG. 2 – D'autres liens entre pages.

Voici à présent un exemple pour lequel on obtient bien une solution non nulle et même, deux solutions incomparables. Il peut être associé à la situation reprise à la figure 2.

$$\begin{cases} s_2 = s_1/4 + s_3 \\ s_3 = s_1/4 + s_2 \\ s_4 = s_1/4 + s_5 \\ s_5 = s_1/4 + s_4. \end{cases}$$

Ici, on trouve par exemple, $s_1 = s_2 = s_3 = 0$, $s_4 = s_5 = 1/2$ ou bien, $s_1 = s_4 = s_5 = 0$, $s_2 = s_3 = 1/2$. D'un point de vue purement mathématique, cela ne pose aucun problème, mais pour construire un classement, devrait-on donner une importance plus grande aux pages 4 et 5 ou aux pages 2 et 3 et pourquoi ?

Perturbation du modèle

L. Page et S. Brin ont alors dû trouver une parade à ce problème. En effet, bien que le système de notre exemple ne possède aucune solution non nulle, ne peut-on quand même pas classer les pages ? Ils ont alors modifié légèrement la matrice $\mathbf{B}$ par deux opérations successives. La première étape consiste à remplacer les colonnes nulles de $\mathbf{B}$ par des colonnes dont tous les éléments sont égaux à $1/n$ (n étant la dimension de la matrice). Dans notre exemple, on a

$$\mathbf{C} = \begin{pmatrix} 0 & 1/5 & 0 & 1/2 & 0 \\ 1/2 & 1/5 & 0 & 0 & 1/3 \\ 1/2 & 1/5 & 0 & 1/2 & 1/3 \\ 0 & 1/5 & 1/2 & 0 & 0 \\ 0 & 1/5 & 1/2 & 0 & 1/3 \end{pmatrix}.$$

Enfin, on construit une matrice $\mathbf{G}$ à partir de $\mathbf{C}$ comme suit. Tout élément de $\mathbf{C}$ est multiplié par $\alpha = 0,85$ et on lui ajoute $(1-\alpha)/n = 0,15/n$ pour obtenir l'élément correspondant de $\mathbf{G}$. Sur notre exemple,

$$\mathbf{G} = 0,85. \begin{pmatrix} 0 & 1/5 & 0 & 1/2 & 0 \\ 1/2 & 1/5 & 0 & 0 & 1/3 \\ 1/2 & 1/5 & 0 & 1/2 & 1/3 \\ 0 & 1/5 & 1/2 & 0 & 0 \\ 0 & 1/5 & 1/2 & 0 & 1/3 \end{pmatrix} + 0,15. \begin{pmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

$$= \begin{pmatrix} 3/100 & 1/5 & 3/100 & 91/200 & 3/100 \\ 91/200 & 1/5 & 3/100 & 3/100 & 47/150 \\ 91/200 & 1/5 & 3/100 & 91/200 & 47/150 \\ 3/100 & 1/5 & 91/200 & 3/100 & 3/100 \\ 3/100 & 1/5 & 91/200 & 3/100 & 47/150 \end{pmatrix}.$$

Le lecteur pourrait objecter que ces manipulations consistant à remplacer $\mathbf{B}$ par $\mathbf{G}$ font qu'on ne respecte plus exactement le modèle initial. C'est exact, mais cet artifice (ou plutôt, ce tour de force) permet en fait d'assurer l'existence et l'unicité de la solution (normalisée) ainsi que son calcul effectif[5] ! On s'est éloigné quelque peu des deux règles **R1** et **R2**, mais on y gagne l'existence et l'unicité de la solution, c'est-à-dire d'un classement.

Le choix de la valeur $\alpha = 0,85$ n'est pas arbitraire et constitue en fait un bon compromis : plus la valeur de α est proche de 1, plus on est proche du modèle initial (si $\alpha = 1$, alors $\mathbf{G} = \mathbf{C}$), mais d'un autre côté pour des raisons de rapidité et de stabilité des calculs réalisés, il vaut mieux ne pas choisir une valeur trop proche de 1.

Les constructions réalisées permettent d'assurer à $\mathbf{G}$ d'être *primitive* (quelle que soit la signification exacte donnée par les mathématiciens à cet adjectif) et *stochastique* (i.e., la somme des éléments de chaque colonne vaut 1 et cette constatation est immédiate, de par la construction même de $\mathbf{B}$ puis $\mathbf{C}$ et $\mathbf{G}$).

Un théorème datant de la première moitié du siècle passé[6], le *théorème de Perron-Frobenius*[7], précise entre autres qu'une matrice primitive $\mathbf{G}$ possède toujours un vecteur $\mathbf{s}$ à composantes réelles et (strictement) positives satisfaisant précisément $\mathbf{G}\mathbf{s} = t\,\mathbf{s}$ (pour un certain $t > 0$ bien défini, mais ce n'est pas l'endroit pour préciser comment un tel t est défini[8]). A un multiple près (ce qui n'est pas étonnant au vu de (3)), ce vecteur est unique. De plus, le fait que la matrice $\mathbf{G}$ soit stochastique entraîne que ce "fameux" t vaut en fait 1. En conclusion, ce théorème découvert bien avant Internet et même avant l'avènement de l'informatique fournit l'existence et l'unicité d'un classement (normalisé) $\mathbf{s}$ satisfaisant

$$\mathbf{G}\mathbf{s} = \mathbf{s}.$$

Ce théorème de Perron-Frobenius possède encore d'autres avantages ! Il stipule aussi que si l'on calcule les puissances successives de la matrice $\mathbf{G}$: $\mathbf{G}$, $\mathbf{G}^2$, $\mathbf{G}^3$, $\mathbf{G}^4$,... cette suite de matrices *converge*[9] vers une matrice limite dont

[5] Ce n'est pas le tout de savoir que la solution existe, si on ne sait pas la calculer, cela ne sert pas à grand chose...

[6] Oskar Perron (1880–1975) et Ferdinand Georg Frobenius (1849–1917), étaient deux mathématiciens allemands.

[7] Pour la petite histoire, Sergeï Brin est le fils d'un mathématicien professionnel, Michael Brin (Université du Maryland), spécialiste des systèmes dynamiques [1]. Il s'agit d'une branche des mathématiques dans laquelle la théorie de Perron-Frobenius est l'un des outils standards. Ceci explique certainement cela.

[8] Cela montre que l'étude de propriétés "évoluées" d'algèbre linéaire est loin d'être dénuée d'intérêt (suivant les versions, la preuve de ce théorème fait près de cinq pages). On pourra aussi remarquer que le théorème en question est "né" près d'un demi-siècle avant Internet. Il ne faut donc jamais préjuger de l'importance de la recherche pure sans application *a priori* immédiate. Ainsi, faire des mathématiques pour le plaisir, leur beauté ou pour faire avancer l'état de nos connaissances peut, et même doit, constituer un but en soi. Nul ne peut prédire l'impact de tels résultats, ceci est l'apanage de la recherche fondamentale.

[9] Ici, inutile d'être précis sur la notion de convergence, intuitivement, le concept est clair. On est en présence de 25 suites numériques réelles, une pour chaque élément de la matrice.

les colonnes sont toutes égales à $\mathbf{s}$ (c'est-à-dire, le vecteur des scores recherché pour classer les pages web). Nous avons calculé numériquement les premières puissances de $\mathbf{G}$:

$$\mathbf{G}^5 = \begin{pmatrix} 0.14721 & 0.13301 & 0.12861 & 0.13367 & 0.13989 \\ 0.18196 & 0.18931 & 0.19412 & 0.18584 & 0.18545 \\ 0.26597 & 0.26067 & 0.26295 & 0.25602 & 0.26253 \\ 0.16641 & 0.17481 & 0.17349 & 0.17906 & 0.17107 \\ 0.23844 & 0.2422 & 0.24084 & 0.2454 & 0.24107 \end{pmatrix}$$

$$\mathbf{G}^{10} = \begin{pmatrix} 0.13568 & 0.13553 & 0.13545 & 0.13559 & 0.13561 \\ 0.18801 & 0.18804 & 0.1881 & 0.18799 & 0.18802 \\ 0.26173 & 0.26161 & 0.26159 & 0.26159 & 0.26166 \\ 0.17304 & 0.17319 & 0.17322 & 0.17319 & 0.17311 \\ 0.24155 & 0.24163 & 0.24164 & 0.24165 & 0.24159 \end{pmatrix}$$

$$\mathbf{G}^{20} = \begin{pmatrix} 0.13556 & 0.13556 & 0.13556 & 0.13556 & 0.13556 \\ 0.18804 & 0.18804 & 0.18804 & 0.18804 & 0.18804 \\ 0.26163 & 0.26163 & 0.26163 & 0.26163 & 0.26163 \\ 0.17316 & 0.17316 & 0.17316 & 0.17316 & 0.17316 \\ 0.24162 & 0.24162 & 0.24162 & 0.24162 & 0.24162 \end{pmatrix}$$

$$\mathbf{G}^{200} = \begin{pmatrix} 0.13556 & 0.13556 & 0.13556 & 0.13556 & 0.13556 \\ 0.18804 & 0.18804 & 0.18804 & 0.18804 & 0.18804 \\ 0.26163 & 0.26163 & 0.26163 & 0.26163 & 0.26163 \\ 0.17316 & 0.17316 & 0.17316 & 0.17316 & 0.17316 \\ 0.24162 & 0.24162 & 0.24162 & 0.24162 & 0.24162 \end{pmatrix}$$

On remarquera qu'avec une précision de cinq chiffres, il n'y a plus ici aucune différence entre $\mathbf{G}^{20}$ et $\mathbf{G}^{200}$.

En conclusion, avec une centaine produits matriciels (voire 200, dans une situation réelle), on obtient un classement plus qu'utile des pages web. Bien sûr, même si effectuer un produit matriciel n'est pas difficile, il faut garder à l'esprit, qu'avec Internet on doit élever à la puissance 200 une matrice de dimension 10^{10}... Ainsi, des super-ordinateurs recalculent 24 heures sur 24 ces puissances sur une matrice constamment remise à jour au gré des nouveaux liens qui se créent ou qui disparaissent sur la toile. Bien évidemment, des développements plus fins permettent d'améliorer certains calculs et on imagine aisément qu'il s'agit d'un secteur suscitant de nombreuses recherches.

5 Interprétation probabiliste

Après les matrices, des probabilités ! En effet, on a parlé de matrices *stochastiques*[10] et pour cause. Le modèle discuté dans les sections précédentes s'interprète de la manière suivante. Imaginez un surfeur passant de page en page à

[10] Voici ce qu'en dit le dictionnaire, `stochastique` : Adjectif

A. – ÉPISTÉMOL. Qui dépend, qui résulte du hasard. *Phénomène stochastique. Poincaré signale dans le détail l'importance, la difficulté, les cas d'exception possibles pour ce problème de Maxwell-Boltzmann. Il fait enfin une allusion précise à ce processus stochastique d'évolution des molécules (Hist. gén. sc., t. 3, vol. 1, 1961, p. 92).*

B. – MATH., STAT. Qui relève du domaine de l'aléatoire, du calcul des probabilités. *Équation, intégrale stochastique. En théorie des probabilités, on dit qu'un phénomène est stochastique s'il dépend de variable(s) aléatoire(s) (Le Garff 1975).*

chaque unité de temps (un gong retentit par exemple, chaque seconde, et à cet instant précis, le surfeur change de page). Se trouvant sur une page donnée p_i, quand le gong sonne, il bascule aléatoirement sur une autre page p_j. On peut alors considérer que l'élément se trouvant en j-ième ligne et i-ième colonne de la matrice $\mathbf{G}$ encode la probabilité, lorsque le surfeur se trouve sur une page p_i, de visiter la fois suivante la page p_j,

$$\mathbb{P}(p_i \to p_j).$$

Voici donc l'interprétation de $\mathbf{G}$ dans le cas d'une matrice 5×5,

$$\mathbf{G} = \begin{pmatrix} \mathbb{P}(p_1 \to p_1) & \mathbb{P}(p_2 \to p_1) & \mathbb{P}(p_3 \to p_1) & \mathbb{P}(p_4 \to p_1) & \mathbb{P}(p_5 \to p_1) \\ \mathbb{P}(p_1 \to p_2) & \mathbb{P}(p_2 \to p_2) & \mathbb{P}(p_3 \to p_2) & \mathbb{P}(p_4 \to p_2) & \mathbb{P}(p_5 \to p_2) \\ \mathbb{P}(p_1 \to p_3) & \mathbb{P}(p_2 \to p_3) & \mathbb{P}(p_3 \to p_3) & \mathbb{P}(p_4 \to p_3) & \mathbb{P}(p_5 \to p_3) \\ \mathbb{P}(p_1 \to p_4) & \mathbb{P}(p_2 \to p_4) & \mathbb{P}(p_3 \to p_4) & \mathbb{P}(p_4 \to p_4) & \mathbb{P}(p_5 \to p_4) \\ \mathbb{P}(p_1 \to p_5) & \mathbb{P}(p_2 \to p_5) & \mathbb{P}(p_3 \to p_5) & \mathbb{P}(p_4 \to p_5) & \mathbb{P}(p_5 \to p_5) \end{pmatrix}$$

Puisque la somme des éléments de chaque colonne vaut 1, cela traduit bien que la probabilité sur l'ensemble des événements possibles (visiter n'importe quelle page à partir d'une page donnée) fait 1. Autrement dit, pour toute page p_i, $i \in \{1, \ldots, n\}$, s'il y a en tout n pages,

$$\mathbb{P}(p_i \to p_1) + \mathbb{P}(p_i \to p_2) + \cdots + \mathbb{P}(p_i \to p_n) = 1.$$

Dans notre exemple, se trouvant sur la page 1, le surfeur visitera la page 2 à l'unité de temps suivante avec une probabilité de $91/200$, car il s'agit de l'élément se trouvant en deuxième ligne et première colonne.

Revenons à la construction de la matrice $\mathbf{G}$ et à son interprétation probabiliste. Se trouvant sur une page donnée, le surfeur a deux options :

a) avec une probabilité de $0,85$, il choisit aléatoirement un des liens présents sur la page actuellement visitée (chacun des liens pouvant être choisi de manière équiprobable, il clique au hasard)

b) avec une probabilité de $0,15$, il est redirigé aléatoirement vers une page quelconque de l'ensemble du web (chacune des pages ayant une même probabilité $1/n$ d'être choisie lors de cette redirection, si n représente le nombre de pages web de l'Internet tout entier).

On peut donc modéliser la situation comme ci-dessous. Sur notre exemple, de chacune des pages, partent cinq arcs portant chacun une pondération qui représente la probabilité que cet arc soit choisi à l'étape suivante en tenant compte de **a)** et **b)** :

$$p_i \overset{\mathbb{P}(p_i \to p_j)}{\longrightarrow} p_j$$

A la figure 3, on n'a pas représenté tous les arcs pour ne pas surcharger celle-ci. En effet, on aurait dû dessiner 5 arcs sortant de chacune des 5 pages. Se trouvant sur une page donnée, le surfeur lance un dé pipé à cinq faces (dé qui tient compte des probabilités correspondantes et ce dé est différent pour chacune des pages). Ensuite, il change de page suivant le résultat du tirage[11].

Que représente $\mathbf{G}^2$ en ces termes probabilistes ? Pour répondre à cette question, il faut repenser à la définition même du produit matriciel. Si l'on désire

[11] Il s'agit en fait des bases des *chaînes de Markov*.

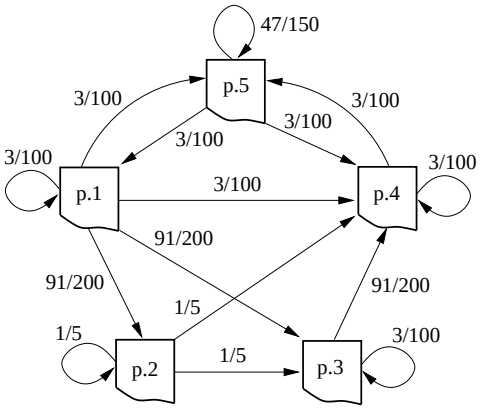

FIG. 3 – Probabilité de transitions.

obtenir l'élément se trouvant à la deuxième ligne et à la quatrième colonne de $\mathbf{G}^2$, on va sommer les produits des éléments correspondants dans ces deux rangées et obtenir :

$$\begin{aligned}
& \mathbf{G}_{21}\mathbf{G}_{14} + \mathbf{G}_{22}\mathbf{G}_{24} + \mathbf{G}_{23}\mathbf{G}_{34} + \mathbf{G}_{24}\mathbf{G}_{44} + \mathbf{G}_{25}\mathbf{G}_{54} \\
= \; & \mathbb{P}(p_1 \to p_2)\mathbb{P}(p_4 \to p_1) + \mathbb{P}(p_2 \to p_2)\mathbb{P}(p_4 \to p_2) + \mathbb{P}(p_3 \to p_2)\mathbb{P}(p_4 \to p_3) \\
& + \mathbb{P}(p_4 \to p_2)\mathbb{P}(p_4 \to p_4) + \mathbb{P}(p_5 \to p_2)\mathbb{P}(p_4 \to p_5) \\
= \; & \mathbb{P}(p_4 \to p_1)\mathbb{P}(p_1 \to p_2) + \mathbb{P}(p_4 \to p_2)\mathbb{P}(p_2 \to p_2) + \mathbb{P}(p_4 \to p_3)\mathbb{P}(p_3 \to p_2) \\
& + \mathbb{P}(p_4 \to p_4)\mathbb{P}(p_4 \to p_2) + \mathbb{P}(p_4 \to p_5)\mathbb{P}(p_5 \to p_2)
\end{aligned}$$

En fait, ce nombre représente exactement la probabilité qu'a le surfeur de se retrouver en deux unités de temps (i.e., avec un "chemin" de longueur 2) à la page 2 s'il est parti de la page 4. En effet, on passe bien en revue tous les chemins de longueur 2 de la page 4 vers la page 2 comme illustré à la Figure 4. Par exemple, le produit $\mathbb{P}(p_4 \to p_1)\mathbb{P}(p_1 \to p_2)$ représente[12] la probabilité d'aller

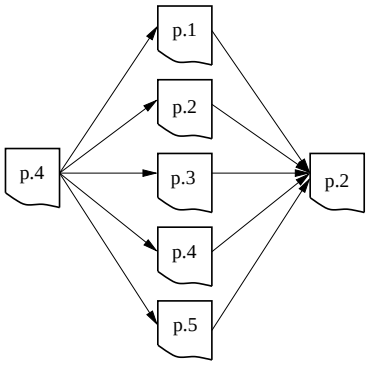

FIG. 4 – Tous les chemins de longueur 2 de la page 4 vers la page 2.

[12]Le lecteur attentif remarquera qu'il s'agit de *processus sans mémoire* : la probabilité de basculer dans une page ne dépend pas du passé (i.e., de l'historique de navigation ou du chemin parcouru précédemment par le surfeur), hypothèse que nous n'avions pas explicitée plus haut.

de la page 4 à la page 2 en passant par la page 1.

Cette constatation est en fait tout à fait générale, la matrice $\mathbf{G}^2$ encode les probabilités de transitions entre pages, en deux coups ! Si on poursuit ce raisonnement, on obtient (par une simple récurrence sur n) que $\mathbf{G}^n$ encode les probabilités de transitions entre pages après n coups. Le théorème de Perron-Frobenius stipule que la limite des matrices $\mathbf{G}^n$ existe quand n tend vers l'infini et donc, cette limite contient les *fréquences*[13] avec lesquelles les différentes pages vont être visitées !

6 Et les équipes de Basket aussi !

La technique développée pour classer des pages peut s'appliquer à d'autres situations comme un championnat sportif. Nous avons choisi le basket (on aurait aussi pu prendre le tennis) mais pas le football, car il n'y a pas de match nul au basket. Ainsi, un match est toujours gagné ou perdu. Si une équipe A bat une équipe B, on trace un arc allant de B vers A dans le graphe symbolisant les différentes rencontres qui ont eu lieu au cours de la saison sportive. Dans un pays comme les États-Unis, le championnat est divisé en deux conférences, celle de l'Est et celle de l'Ouest. Les équipes d'une même conférence jouent entre elles et puis, certains matchs sont organisés entre les deux conférences (mais il y a moins de matchs de ce dernier type, tout le monde ne joue pas contre tout le monde, alors que toutes les équipes d'une même conférence se sont au moins rencontrées une fois). Si une équipe gagne tous les matchs de la conférence Est, cela signifie-t-il qu'elle est la meilleure du pays, si, dans le même temps, aucune équipe de la conférence Ouest n'est invaincue ? Certainement pas. Il se peut que le championnat de la conférence Ouest soit bien plus disputé et relevé car s'y présentent plusieurs équipes de niveau comparable, alors que dans la conférence Est, les équipes sont très moyennes sauf une qui les surclasse. Cette équipe n'occuperait d'ailleurs peut-être que le milieu de classement, si elle jouait dans la conférence Ouest. Les deux règles **R1** et **R2** s'adaptent alors parfaitement ici.

R1 On accorde d'autant plus d'importance aux matchs gagnés contre une équipe réputée forte, c'est-à-dire qui possède un score élevé.

R2 On accorde d'autant moins d'importance aux matchs gagnés contre une équipe qui perd toutes ses rencontres ou presque.

Il suffit alors de reprendre exactement la même machinerie que précédemment. On remplace simplement les pages par des équipes et les liens par une liaison symbolisant une victoire.

[13] Nous nous autoriserons une fois encore à nous baser sur l'intuition...

7 Appendice sur le calcul matriciel

Nous rappelons ici, au moyen de quelques exemples, comment multiplier une matrice par un vecteur, deux matrices ou encore comment calculer la puissance n-ième d'une matrice.

Commençons par un exemple, pour multiplier une matrice 3×3 par un vecteur colonne contenant 3 éléments, on procède comme suit

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{pmatrix}.$$

Ainsi, pour trouver une composante du vecteur résultant, il suffit de faire la somme des produits des éléments de la ligne correspondante de la matrice par les éléments correspondants du vecteur. Cette règle s'étend à une matrice $n \times n$ et à un vecteur à n composantes.

Pour multiplier deux matrices, on procède essentiellement de la même manière. En effet, une matrice 3×3 peut être vue comme la juxtaposition de 3 vecteurs formant ces colonnes. Puisque nous savons à présent multiplier une matrice par un vecteur, il suffit de répéter l'opération 3 fois, pour chacune des colonnes de la deuxième matrice. Il vient

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}$$

$$= \begin{pmatrix} ax_1 + by_1 + cz_1 & ax_2 + by_2 + cz_2 & ax_3 + by_3 + cz_3 \\ dx_1 + ey_1 + fz_1 & dx_2 + ey_2 + fz_2 & dx_3 + ey_3 + fz_3 \\ gx_1 + hy_1 + iz_1 & gx_2 + hy_2 + iz_2 & gx_3 + hy_3 + iz_3 \end{pmatrix}.$$

Enfin, le carré d'une matrice n'est rien d'autre que le produit de la matrice par elle-même. Il s'agit donc d'un cas particulier de multiplication de deux matrices carrées de même dimension. De même, on obtient, de proche en proche, la puissance n-ième d'une matrice $\mathbf{A}$ en remarquant que $\mathbf{A}^{n+1} = \mathbf{A}.\mathbf{A}^n$.

Références

[1] M. Brin, G. Stuck, Introduction to Dynamical Systems, Cambridge University Press, 2002.

[2] P. Fernández Gallardo, Google's secret and Linear Algebra, *EMS (European Mathematical Society) Newsletter March 2007*, `http ://www.uam.es/personal_pdi/ciencias/gallardo/ems63-pablo-fernandez_final.pdf`

[3] C. Meyer, A. Langville, Google's PageRank and Beyond : The Science of Search Engine Rankings, Princeton University Press, (1996).

[4] L. Page, S. Brin, R. Motwani, T.Winograd, *The PageRank Citation Ranking : Bringing Order to the Web*, Technical Report, Stanford University, 1998.

[5] J. Miles Prystowsky, L. Gill, *Calculating Web Page Authority Using the PageRank Algorithm*,
`http ://online.redwoods.cc.ca.us/instruct/darnold/LAPROJ/`.

[6] M. Rigo, *Théorie des graphes*, notes de cours, 2ièmes bacheliers en sciences mathématiques, Université de Liège, 2007–2008.
`http ://www.discmath.ulg.ac.be/`

[7] E. Seneta, *Non-Negative Matrices, An Introduction to Theory and Applications*, George Allen and Unwin Ltd, London, (1973).

La référence [1] est un livre écrit par le père de S. Brin. Il ne contient qu'un court passage sur Google (4 pages) et est axé sur les systèmes dynamiques. Les références [2, 5] sont des documents assez généraux expliquant le fonctionnement de Google (ils s'adressent néanmoins assez rapidement à un lecteur ayant déjà des bases solides d'algèbre). Pour une étude en profondeur, [3] est LA référence. On y traite de manière exhaustive les problèmes, les techniques et les solutions liés à ce moteur de recherche (avec une annexe mathématique très détaillée, discussion des valeurs propres de Google, etc. . .). L'article [4] contient les idées fondatrices de Brin et Page. Y apparaissent non seulement les idées sur le "PageRank" mais il traite aussi des autres aspects du moteur de recherche (bases de données, extraction de données, etc. . .). Enfin, [7] traite dans son premier chapitre du théorème de Perron-Frobenius en détails et [6] contient une section reprenant les développements omis dans le présent document (avec par exemple, un énoncé détaillé du théorème de Perron-Frobenius).

Michel Rigo
Département de Mathématiques
Université de Liège
Grande Traverse 12 (B37)
4000 Liège
`M.Rigo@ulg.ac.be`
`http ://www.discmath.ulg.ac.be/`