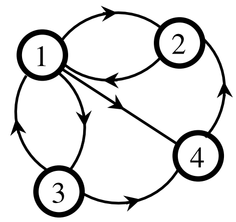
Comment Google classe les pages web

Niveau : terminale générale, Maths Expertes.

Lien avec le programme : écriture matricielle d'un système linéaire, suite de matrices lignes vérifiant une relation de récurrence du type $U_{n+1} = U_n A + C$, recherche d'une suite constante vérifiant la relation de récurrence, étude de la convergence. Chaine de Markov.

Prérequis : matrice, opérations sur les matrices, raisonnement par récurrence, probabilités conditionnelles, limite de la suite réelle (q^n).

Lien avec *Les maths au quotidien* : **Navigation**.



Un internaute (fictif) ayant tapé des mots-clés très spécialisés dans son moteur de recherche arrive sur les quatre pages correspondantes sur le web. Elles sont reliées entre elles par des liens hypertexte comme sur le graphe suivant :
On note $i \to j$ un lien de la page i vers la page j.
À chaque clic, l'internaute change de page.
On va étudier la position du surfeur au bout de n clics.
Pour cela, on considère la variable aléatoire X_n égale respectivement à 1, 2, 3 ou 4 selon que l'internaute est sur la page 1, 2, 3 ou 4 au $n$ième clic.

Partie A Premier modèle

L'internaute change de page en cliquant « au hasard » sur un des liens de la page où il se trouve, sans préférence particulière pour tel ou tel lien.

1. On suppose que l'internaute part de la page 1.
On a donc $P(X_0 = 1) = 1$, $P(X_0 = 2) = 0$, $P(X_0 = 3) = 0$, $P(X_0 = 4) = 0$.

a. Déterminer $P(X_1 = 1)$, $P(X_1 = 2)$, $P(X_1 = 3)$, $P(X_1 = 4)$.

b. Déterminer $P(X_2 = 1)$, $P(X_2 = 2)$, $P(X_2 = 3)$, $P(X_2 = 4)$. On pourra faire un arbre de probabilités.

2. L'internaute commence sa navigation sur une page quelconque.

a. Montrer que $P(X_{n+1} = 1) = P(X_n = 2) + \frac{1}{2} P(X_n = 3)$.

b. Exprimer de la même manière $P(X_{n+1} = 2)$, $P(X_{n+1} = 3)$ et $P(X_{n+1} = 4)$ en fonction de $P(X_n = 1)$, $P(X_n = 2)$, $P(X_n = 3)$ et $P(X_n = 4)$.

c. Soit U_n la matrice ligne $(P(X_n = 1) \;\; P(X_n = 2) \;\; P(X_n = 3) \;\; P(X_n = 4))$.
Déterminer la matrice A telle que $U_{n+1} = U_n A$.
A est appelée la matrice de transition. Que représente le coefficient a_{ij} de cette matrice ?

d. En déduire l'expression de U_n en fonction de U_0 et de n.

e. Déterminer les probabilités pour l'internaute d'être sur les différentes pages au bout de 20 étapes si celui-ci part de la page 1. Mêmes questions si celui-ci part de la page 2, de la page 3, de la page 4.

f. À l'aide de la calculatrice ou d'un logiciel, calculer A^{10}, A^{80}, A^{150}, A^{250} (on arrondira les coefficients des matrices à 10^{-4} près).
Émettre une conjecture sur l'éventuelle matrice limite B de la suite de matrice (A^n) et en déduire les probabilités asymptotiques pour l'internaute d'être sur les différentes pages web.

g. Déterminer la matrice ligne X, dont la somme des coefficients est égale à 1, et telle que $X A = X$.
Que remarque-ton ?

Partie B Deuxième modèle

En réalité, il existe beaucoup de pages web qui ne possèdent pas de lien hypertexte. Par conséquent si l'on considère qu'un internaute change de page seulement par lien hypertexte, et si celui-ci arrive sur l'une de ces pages, il se retrouvera bloqué. De manière plus subtile, il peut aussi arriver que l'internaute soit bloqué dans une portion du web, sans pouvoir s'y échapper. De plus, chacun sait que l'on peut choisir à tout moment une autre page donnée par le moteur de recherche et ceci sans suivre de lien donné.
C'est pourquoi Google utilise un modèle plus raffiné, et l'internaute a la possibilité de s'*échapper* et de revenir sur l'une quelconque des pages qui traitent son sujet.

Reprenons notre exemple précédent, mais en introduisant un paramètre $c \in [0\,;\,1]$, appelé *coefficient d'échappement* :
– avec probabilité c, le surfeur abandonne la page actuelle et recommence sur une des 4 pages du web, choisie de manière équiprobable (il peut rester sur la page actuelle…).
– avec probabilité $1 - c$, le surfeur suit un des liens de la page actuelle i, choisi de manière équiprobable parmi tous les liens proposés par cette page.

Remarque : la valeur $\frac{1}{c}$ est le nombre moyen de pages visitées avant de recommencer aléatoirement sur une page.

Comme les concepteurs de Google, on va prendre $c = 0{,}15$, ce qui correspond à suivre environ 6 liens en moyenne.

Encore une fois, on va étudier la position du surfeur au bout de n clics.
Pour cela, on note à nouveau X_n la variable aléatoire égale respectivement à 1, 2, 3 ou 4 selon que l'internaute est sur la page 1, 2, 3 ou 4 après le $n$ième clic. A est la matrice de la partie A donnant les probabilités de passage d'une page à une autre **en suivant un lien** et on note à nouveau U_n la matrice ligne $(P(X_n = 1)\ \ P(X_n = 2)\ \ P(X_n = 3)\ \ P(X_n = 4))$.
On note C la matrice ligne dont les quatre coefficients sont 0,25.

1. Justifier que la suite de matrices lignes (U_n) vérifie la relation $U_{n+1} = U_n\,(0{,}85A) + 0{,}15C$.

2. On cherche une matrice ligne L telle que $L = L\,(0{,}85A) + 0{,}15C$.

a. Montrer que $L\,(\mathrm{I_d} - 0{,}85A) = 0{,}15C$.

b. En déduire que si $(\mathrm{I_d} - 0{,}85A)$ est inversible, L existe et de manière unique.

c. Vérifier avec un logiciel ou la calculatrice que $(\mathrm{I_d} - 0{,}85A)$ est inversible et déterminer son inverse. Déterminer alors la matrice ligne L.

3. On pose, pour tout nombre entier naturel n, $V_n = U_n - L$.

a. Démontrer que pour tout nombre entier naturel n, $V_{n+1} = V_n\,(0{,}85A)$.

b. En déduire que, pour tout nombre entier naturel n, $V_n = V_0\,(0{,}85^n A^n)$.

c. Si l'on admet la conjecture réalisée dans la première partie quant à la limite de la suite (A^n), conclure sur la limite de (V_n) puis sur celle de (U_n).

Point info : les pertinences des pages 1, 2, 3, 4 sont données par les coefficients respectifs de la matrice ligne L.