Pertinence d'une page web

Ressources pour la classe terminale
générale et technologique

QUELQUES PASSAGES
DE LA PREMIERE PARTIE

Pertinence d'une page web

Mathématiques
Série S

Enseignement de spécialité

1. De la recherche dans une bibliothèque à la recherche dans un graphe

Un moteur de recherche doit fournir à chaque utilisateur une liste de pages où apparaissent des mots-clés donnés dans la requête de celui-ci. On peut avoir l'idée de classer les milliards de pages disponibles dans un ordre permettant le tri à partir des mots-clés fournis. Cela demande des moyens de stockage considérables et la réorganisation continuelle (en temps réel, comme on dit) de ces archives. Il faut de plus assurer aux milliers de requêtes simultanées des réponses rapides, mais aussi des réponses fiables.

Un moteur de recherche copie dans un premier temps les pages web sur des milliers d'ordinateurs et les trie par ordre alphabétique des mots clés. La première idée simple consisterait pour chaque requête à fournir la liste de pages contenant le (ou les) mots clés de la requête. Mais il y en a des dizaines de milliers ! Aussi l'ordre alphabétique n'apparaît pas le meilleur pour assurer un service rapide et de qualité. Les pages référencées pour le client doivent donner une idée aussi juste que possible de l'information disponible au moment de la requête et faire apparaître en premières citations celles qui y répondent le mieux, les plus *pertinentes*.

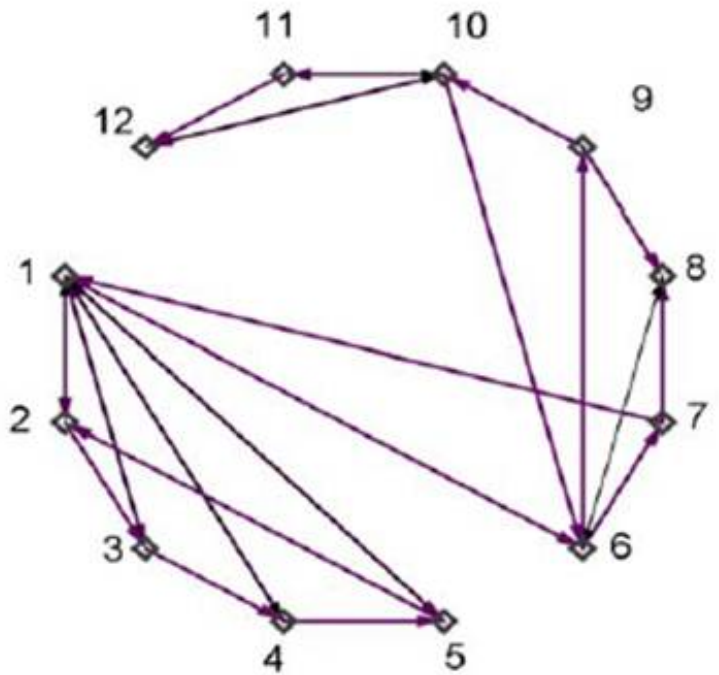
Le web n'est pas une simple bibliothèque de pages web. Les pages web comportent des *liens* qui permettent d'accéder directement de l'une à d'autres. On peut donc considérer le web comme un graphe orienté, dont chaque page web est un sommet et chaque lien est un arc. L'idée pour déterminer la pertinence d'une page en lien avec un mot clé va être de s'appuyer sur l'existence de ces liens, en partant de l'idée basique que plus une page est citée, plus elle est *pertinente*.

Dans la suite, les pages web sont numérotées 1, 2, …, i, …, n et un lien de la page i vers la page j est noté $i \to j$.

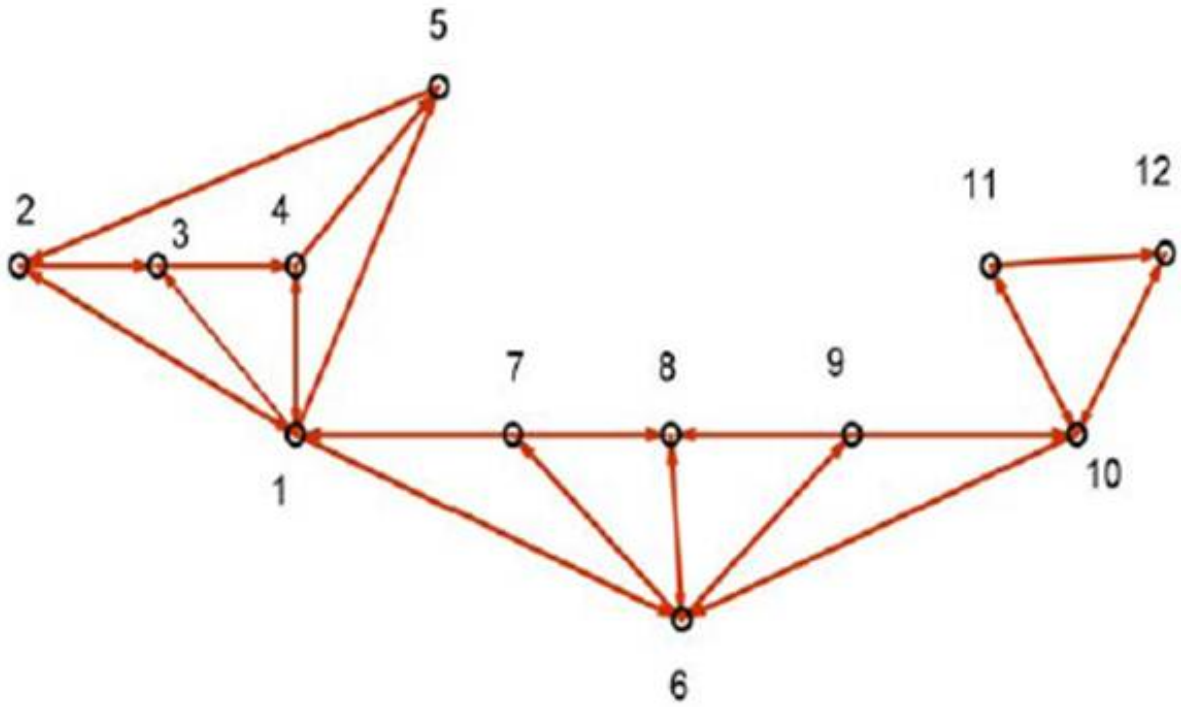
Ainsi on cherche à attribuer à chacune des pages une mesure de pertinence (un nombre réel ≥ 0).

2. Un exemple

Pour la suite, nous allons considérer un exemple excessivement simple avec seulement 12 pages web et les liens suivants :



Le graphe ci-dessus, qui ne comporte pourtant que 12 sommets, n'est pas très lisible. Les sommets les plus « fréquentés » n'y sont pas facilement identifiables.



Une nouvelle représentation de ce graphe, plus « buissonnante » à défaut d'être arborescente, met mieux en évidence l'importance des sommets 1, 6 et 10, vers lesquels « pointent » un nombre élevé d'autres sommets.

3. Mesurer la pertinence

Dans ce qui suit, on note μ_j la mesure de pertinence de la page j, pour tout entier j compris entre 1 et n, nombre de pages web disponibles à l'instant considéré, en rapport avec la requête considérée.

- Comptage naturel des liens

A chaque page j, on associe le nombre de liens $i \rightarrow j$ qui pointent vers elle.

Dans notre exemple, les pages 6 et 10 reçoivent chacune 3 liens, tandis que la page 1 en reçoit 5. On obtient donc : $\mu_6 = 3$, $\mu_{10} = 3$ et $\mu_1 = 5$.

Mais ce comptage n'est pas suffisamment discriminant et il est de plus très facile à manipuler, puisqu'il suffit de créer des « fausses » pages pointant vers la page i pour en augmenter l'importance.

- Comptage pondéré

On peut tenter de pondérer les liens : certaines pages émettent beaucoup de liens ce qui d'une certaine façon diminue leur poids.

Notons, pour chaque entier i, λ_i le nombre de liens émis par la page i.

On peut alors définir la mesure de pertinence de la page j en comptant le nombre de liens pondérés qui pointent vers elle :

$$\mu_j = \sum_{i \to j} \frac{1}{\lambda_i}$$

Dans notre exemple, $\mu_6 = 1{,}45$, $\mu_{10} = 1{,}5$ et $\mu_1 = 2{,}5$.

Mais cette mesure présente toujours le même risque d'être manipulée.

- Comptage récursif

La pertinence d'une page est renforcée par la pertinence des pages qui pointent vers elle et elle est diminuée par la dispersion éventuelle des liens issus de ces dernières.

En reprenant la pondération précédente, on peut définir la pertinence d'une page *j* de la façon suivante :

$$\mu_j = \sum_{i \to j} \frac{1}{\lambda_i} \mu_i \quad (*)$$

Le risque de manipulation consistant en l'ajout de pages vides de sens est ici annulé puisqu'une telle page recevrait une mesure de pertinence nulle.

Avec le graphe présenté dans le paragraphe précédent, on obtient par exemple :

$$\mu_7 = \frac{1}{3}\mu_6, \quad \mu_{12} = \mu_{11} + \frac{1}{2}\mu_{10}, \text{ etc.}$$

On obtient ainsi un système d'équations linéaires.

Notons, pour chaque entier i, λ_i le nombre de liens émis par la page i.

On réécrit les formules (*) pour tout entier i compris entre 1 et n avec des coefficients notés $a_{i\,j}$, le coefficient $a_{i\,j}$ valant $\frac{1}{\lambda_i}$ si la page i pointe vers la page j, 0 sinon. On obtient ainsi le système linéaire de n équations à n inconnues (les μ_i). :

$$\mu_j = \sum_{i=1}^{n} a_{ij}\mu_i \qquad 1 \leq j \leq n$$

Les coefficients $a_{i\,j}$ définissent une matrice à n lignes et n colonnes (de format (n, n)), que l'on peut noter A.

Le système linéaire de n équations précédent correspond à l'équation matricielle suivante :

$$W = WA,$$

où W est une matrice ligne à n colonnes (format $(1, n)$), dont les coefficients sont les μ_j :

$$W = (\mu_1 \ \mu_2 \ \ldots \ \mu_n)$$

« Il n'y a plus qu'à » résoudre ce système, sauf que dans le cas du web il y a des milliards d'inconnues. Dans notre exemple, on obtient une matrice de format (12,12).

4. Pertinence et probabilités

Dans le système d'équations ($\mu_j = \sum_{i=1}^{n} a_{ij}\mu_i$, $1 \leq j \leq n$) précédent, on peut remarquer la propriété suivante des coefficients a_{ij} :

$$\sum_{j=1}^{n} a_{ij} = 1$$

En fait, pour un indice i fixé (c'est-à-dire dans la ligne i de la matrice) tous les coefficients non nuls sont égaux à l'inverse du nombre de liens émis par la page i, ce nombre correspondant également de ce fait à l'inverse du nombre de coefficients non nuls de la ligne i.

Les coefficients a_{ij} (tous positifs ou nuls) peuvent donc s'interpréter comme la probabilité, pour un « surfeur » qui se trouverait à la page i de suivre le lien qui l'amènerait à la page j. Cette probabilité est définie de la manière suivante : si λ_i liens sont issus de la page i, la probabilité pour que le surfeur aléatoire du web passe de la page i à une des pages vers lesquelles elle pointe est $\frac{1}{\lambda_i}$, la probabilité pour qu'il se dirige vers une autre est 0.

Notons X_p la variable aléatoire indiquant la position (numéro de page) du surfeur aléatoire après p clics. On a :

$$P(X_{p+1} = j) = \sum_{i=1}^{n} P_{[X_p=i]}(X_{p+1} = j).P(X_p = i) = \sum_{i=1}^{n} a_{ij}P(X_p = i)$$

En notant U_p la matrice ligne à n colonnes admettant $P(X_p = i)$ pour coefficient à la colonne i pour tout entier i compris entre 1 et n, les relations précédentes peuvent se traduire par la relation matricielle suivante :

$$U_{p+1} = U_p A$$

On en déduit par récurrence que, pour tout entier p strictement positif, $U_p = U_0 A^p$, avec U_0 donnant la position du surfeur aléatoire au départ (U_0 est donc une matrice ligne à n éléments tous nuls sauf un qui vaut 1 et dont l'indice correspond au numéro de la page de départ).

Toutefois, il peut arriver que certaines pages ne comportent aucun lien vers d'autres pages ; dans ce cas, lorsque le surfeur aléatoire arrive sur l'une d'entre elles, il lui est impossible de la quitter. La ligne de la matrice correspondant à cette page ne comporte alors que des 0. Afin de remédier à ce défaut et sans doute coller mieux à la réalité, on introduit la possibilité de quitter à tout instant une page quelconque pour se diriger vers une autre choisie au hasard, et ce avec une probabilité égale à c.

Dans ces conditions, le modèle correspond au système de relations suivant pour tout entier p strictement positif et tout entier i compris entre 1 et n (puisque $\sum_{j=1}^{j=n} P(X_p = j) = 1$) :

$$P(X_{p+1} = j) = \frac{c}{n} + (1-c)\sum_{i=1}^{n} P_{[X_p=i]}(X_{p+1} = j).P(X_p = i) = \sum_{i=1}^{n} (\frac{c}{n} + (1-c)a_{ij}).P(X_p = i)$$

qui se traduit par la relation matricielle suivante (pour tout entier $p > 0$) :

$$U_{p+1} = U_p\,[\frac{c}{n}J + (1-c)A]$$

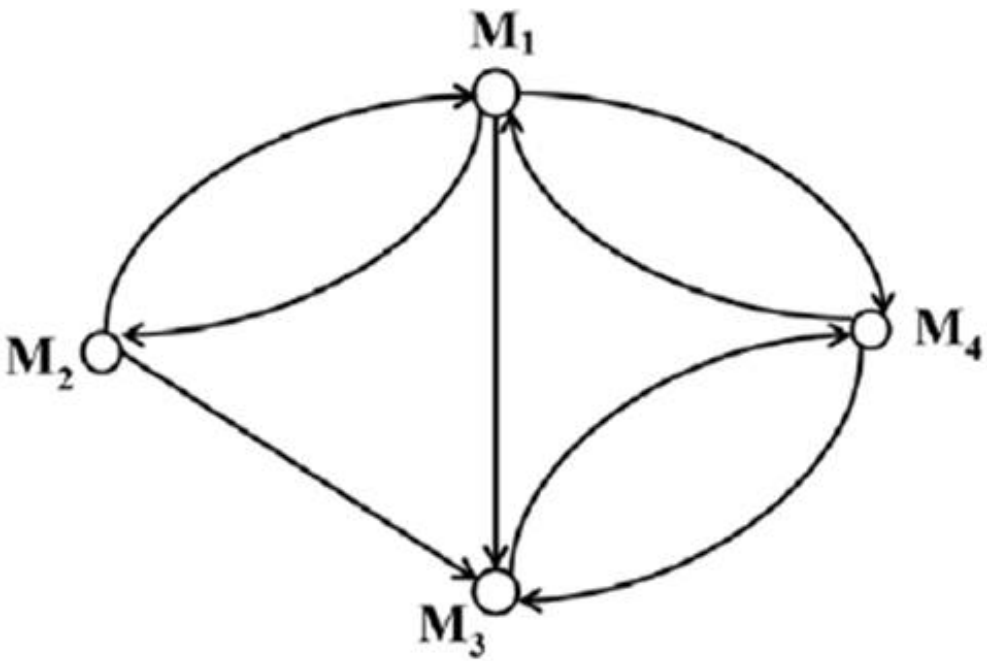
où J désigne la matrice carrée de format (n, n) dont tous les coefficients sont égaux à 1. Notons $B = \frac{c}{n}J + (1-c)A$.

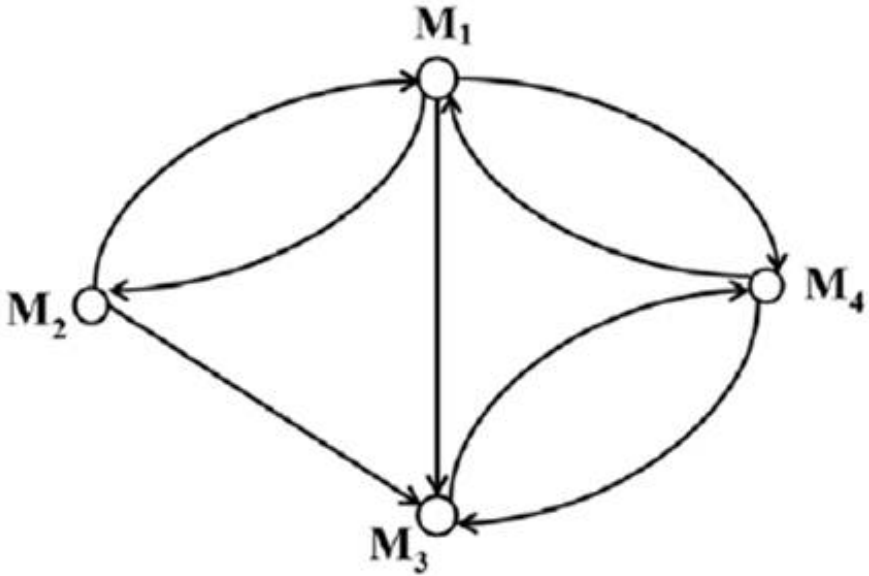
On a alors, pour tout entier p strictement positif, $U_p = U_0 B^p$.

Il resterait à prouver que la suite de matrice (B^p) converge (dans un sens à préciser) lorsque l'entier p tend vers l'infini et expliquer comment récupérer les mesures de pertinence $\mu_1, \mu_2, \ldots, \mu_n$.

Il n'est pas question de traiter ici le cas général. Nous allons nous contenter d'observer ce qui se passe sur un exemple élémentaire.

Dans l'exemple ci-dessous, le graphe représente les liens existant entre quatre pages web numérotées de 1 à 4 (M_1, M_2, M_3, M_4) :





La matrice A associée à ce graphe, telle que définie dans le paragraphe précédent, est :

$$A = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

On observe ici que la matrice n'est pas symétrique contrairement à la matrice d'adjacence définie dans la partie *d*. Cela tient au fait qu'ici le graphe est orienté.

On peut montrer que les puissances de la matrice A ont pour limite la matrice L ci-dessous. On a alors, quelle que soit la situation initiale U_0, $U_0L = \left(\frac{3}{13} \quad \frac{1}{13} \quad \frac{4}{13} \quad \frac{5}{13}\right)$.

En conséquence, on attribue aux pages 1, 2, 3, 4 les indices de pertinence respectifs $\frac{3}{13}$, $\frac{1}{13}$, $\frac{4}{13}$ et $\frac{5}{13}$.

$$L = \begin{pmatrix} \frac{3}{13} & \frac{1}{13} & \frac{4}{13} & \frac{5}{13} \\ \frac{3}{13} & \frac{1}{13} & \frac{4}{13} & \frac{5}{13} \\ \frac{3}{13} & \frac{1}{13} & \frac{4}{13} & \frac{5}{13} \\ \frac{3}{13} & \frac{1}{13} & \frac{4}{13} & \frac{5}{13} \end{pmatrix}$$

Dans le deuxième modèle, avec $c = 1 / 5$, on obtient la matrice :

$$B = \frac{1}{20}J + \frac{4}{5}A = \begin{pmatrix} \frac{1}{20} & \frac{19}{60} & \frac{19}{60} & \frac{19}{60} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{17}{20} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \end{pmatrix}$$

On démontre que les puissances de la matrice B conduisent à une matrice limite et à des indices de pertinence qui sont $\frac{135}{572}$, $\frac{323}{2860}$, $\frac{171}{572}$ et $\frac{1007}{2860}$, à comparer aux indices trouvés précédemment.

Page	1	2	3	4
Sans saut aléatoire	0,23	0,08	0,31	0,38
Avec saut aléatoire	0,24	0,11	0,3	0,35