Un enseignement au numérique pour notre académie : Principe de fonctionnent d'un moteur de recherche

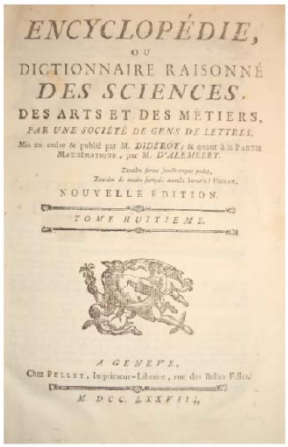
De l'Encyclopédie aux moteurs de recherche

Marie-Danièle CAMPION
Recteur de l'académie de Clermont-Ferrand

La formation au numérique sur notre académie se fixe comme objectif de faire découvrir aux professeurs les aspects conceptuels du numérique. Il s'agit de mieux faire percevoir le monde qui les entoure à nos élèves, en s'appuyant sur les sciences du numérique.
Après le premier texte diffusé à la rentrée 2013 qui sensibilisait les professeurs aux questions liées au numérique, nous souhaitons enrichir les connaissances des personnels en ce qui concerne les moteurs de recherche en s'attachant à décortiquer ici le principe de fonctionnement d'un moteur de recherche.

Que pourraient penser Diderot et d'Alembert des moteurs de recherche fournissant des donnés dont l'intérêt intellectuel n'est pas toujours filtré ? Ces savants du 18e siècle s'étaient donnés pour mission au sein d'une société de gens de lettres de fournir en 28 volumes « l'essentiel » de la culture en des temps où l'école de la République n'existait pas.

Portrait par Van Loo de Denis Diderot 1713-1784


ENCYCLOPÉDIE,
OU
DICTIONNAIRE RAISONNÉ
DES SCIENCES,
DES ARTS ET DES METIERS,
PAR UNE SOCIÉTÉ DE GENS DE LETTRES.
NOUVELLE ÉDITION.
TOME HUITIEME.
A GENEVE,
M. DCC. LXXVIII.

L'Encyclopédie a certainement été quelque part créatrice d'une habitude que nous avons prise de chercher dans l'ordre alphabétique tel ou tel mot afin d'accéder à une liste de pages. Le simple fait de lancer aujourd'hui une recherche sur un mot avec un moteur de recherche nous interroge sur le processus mis en œuvre pour fournir les pages obtenues qui n'a rien à voir avec l'ordre alphabétique.

Page de garde d'un tome de l'Encyclopédie

Les moteurs de recherche sont aujourd'hui au cœur de la recherche documentaire. Mais l'accès à la connaissance n'est pas aussi simple qu'on pourrait l'espérer. Seule une analyse experte des résultats fournis par ces moteurs de recherche peut permettre leur exploitation.

Avec ce document, nous invitons à découvrir une partie du secret de fonctionnement d'un moteur de recherche, secret dont nous pouvons dire qu'il est non seulement bien gardé mais aussi en perpétuelle évolution. Ces moteurs de recherche sont des outils incontournables de la transmission des savoirs, c'est en montrant à l'élève pourquoi Google est différent de l'index de l'encyclopédie du centre de connaissances et de culture que nous arriverons à développer le sens majeur d'une utilisation raisonnée de ces outils.

Nous souhaitons in fine fournir des éléments non seulement intéressants parce qu'ils font partie de ce que l'on appelle la culture numérique mais aussi incontournables dans la formation de nos élèves en ce qui concerne l'analyse des données fournies par ces moteurs de recherche.

Les moteurs de recherche

David Fayon
Expert en technologies numériques
Auteur du livre intitulé « web 2.0 et au-delà »

Monsieur FAYON est le concepteur d'un site entièrement dédié à l'actualité du web et du numérique que vous trouverez à l'adresse suivante : www.david.fayon.fr .

1. Principes de fonctionnement

Un moteur de recherche est un outil sur Internet qui permet de trouver des informations (pages, images, vidéos) associées à des mots saisis par l'internaute. Les informations présentes sur le Web sont de plus en plus nombreuses. De surcroît, elles évoluent en permanence (mise à jour des pages, modification de l'adresse d'hébergement d'un site). Aussi il est apparu nécessaire de développer des moteurs de recherche et des annuaires de sites qui classent les sites par thèmes pour faciliter le travail de recherche de l'internaute.

Concrètement un moteur de recherche comprend deux grandes fonctionnalités.

D'une part l'indexation. Il s'agit pour le moteur de recherche de parcourir les pages du web *via* un robot d'indexation qui va de lien en lien sur les sites (adresse de type http:// ou www) pour indexer les ressources récupérées dans des bases de données. Le robot regarde les modifications et les transmet à un indexeur qui référence le contenu des pages et enregistre le résultat dans un index. Chaque moteur possède son propre algorithme qui, par ailleurs évolutif.

Ceci est effectué à l'image de l'indexation d'un livre qui permet ensuite de retrouver facilement et par ordre alphabétique à quelle page se situe l'information cherchée (par exemple un personnage historique cité dans l'ouvrage).

D'autre part, la recherche des informations indexées par les moteurs (1), phase préalable à la consultation (2). L'internaute va saisir des requêtes (phrase ou mots clés recherchés avec éventuellement des guillemets et des opérateurs logiques) dans un champ de saisie. Le moteur de recherche va faire appel à un algorithme qui va utiliser l'index et trier et présenter les résultats du plus pertinent au moins.

Notons que les moteurs de recherche peuvent disposer de modules complémentaires comme :

- un correcteur orthographique pour déceler d'éventuelles erreurs de saisie,
- un lemmatiseur qui permet de restreindre des mots recherchés à leur forme de référence (par rapport aux pluriels, conjugaisons, etc.) et ne conserver qu'une forme unique,
- un outil de suppression des mots vides (tant dans l'index que dans les requêtes), c'est-à-dire les mots de transition qui ne sont pas indexés dans les bases de données (le, la, de, du, ce, etc.) de façon à augmenter la pertinence des résultats.

2. Les moteurs de recherche aujourd'hui

90 % des recherches en France sont faites avec le moteur de recherche Google. Nous avons également des outils qui proposent des fonctions analogues comme Bing, Yahoo!, Exalead. Aussi il est important que son site ou son blog soit bien vu de Google et de connaître les techniques pour maximiser sa visibilité et apparaître, pour une recherche sur des mots clés donnés en rapport avec son site ou son blog, dans les premiers résultats délivrés par Google.

Globalement, les résultats pour une recherche sont la résultante de la pertinence par rapport aux mots clés saisis par le PageRank[1] d'une site, score compris entre 0 et 10 et qui est lié notamment aux nombres de liens (et à l'importance de ceux-ci) qui pointent vers la page considérée de son site. Les techniques ont été raffinés et la notion de confiance (TrustRank) est également prise en compte dans le résultat.

Résultat pour une requête sur Google =

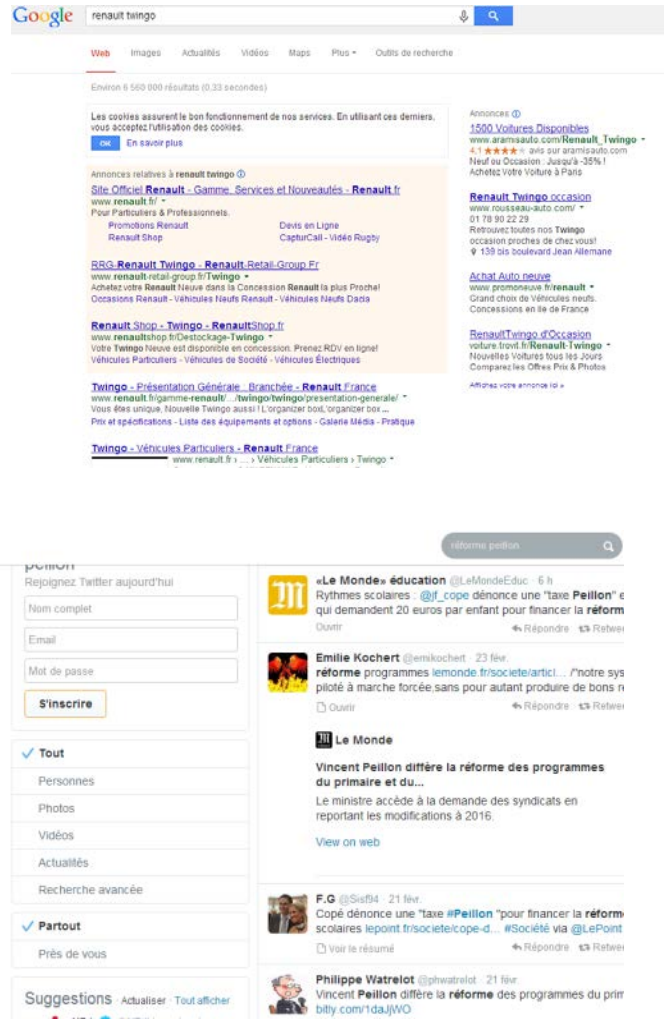
[1] On pourra se référer à www.pagerank.fr

pertinence x PageRank

Une requête donne plusieurs résultats classés par ordre de pertinence décroissant. Les internautes choisissent de cliquer sur les liens qui apparaissent dans les premières positions. D'où la course à l'optimisation pour que les pages de son site soient visibles.

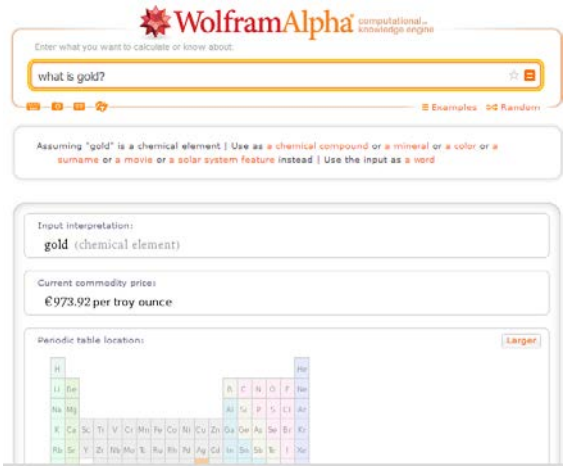
Il est à noter que globalement 70 % des visites sur les sites proviennent du référencement naturel, c'est-à-dire du travail effectué par le webmestre pour développer et optimiser son site (choix des mots clés, titre des pages, adresses des pages URL avec les mots clés en rapport avec le contenu, balises de titres <h1> à <h6>, caractères gras, italiques, etc. pour mettre en exergue certains mots, liens entrants, noms des liens et des images, etc.). Les 30 % restant sont le résultat du référencement payant (achat des mots clés *via* un mécanisme d'enchères, programme AdWords et AdSense de Google).

Typiquement, lors d'une recherche faite sur Google, par exemple « renault twingo », nous avons dans les premiers résultats des liens sur fond rose qui correspondent à des liens dits « sponsorisés » et qui correspondent à l'achat de mots clés. Dès que l'on clique dessus, on se rend à l'espace souhaité et dans l'exemple, Renault verse à Google quelques centimes d'euros pour l'apport de visiteur drainé. Ensuite apparaissent les liens qui proviennent du référencement naturel.



Deux évolutions se dessinent pour les moteurs de recherche. D'abord une évolution vers les moteurs de recherche sémantique qui analysent les requêtes formulées et tentent de donner des réponses davantage en rapport. L'exemple type est Wolfram Alpha, outil disponible pour l'heure en anglais et qui est un moteur de recherche encyclopédique. Pour une requête donnée, une seule réponse est délivrée.

Ensuite, le développement d'autres outils qui permettent de donner des résultats dans l'instantanéité. Ainsi Twitter comprend une zone de recherche et en saisissant un mot clé ou un groupe de mots clés en rapport avec une actualité, des résultats seront donnés par ordre antéchronologique avec, le cas échéant, des tweets qui comprennent des liens vers des sites où figurent une information plus complète, des photos, etc.



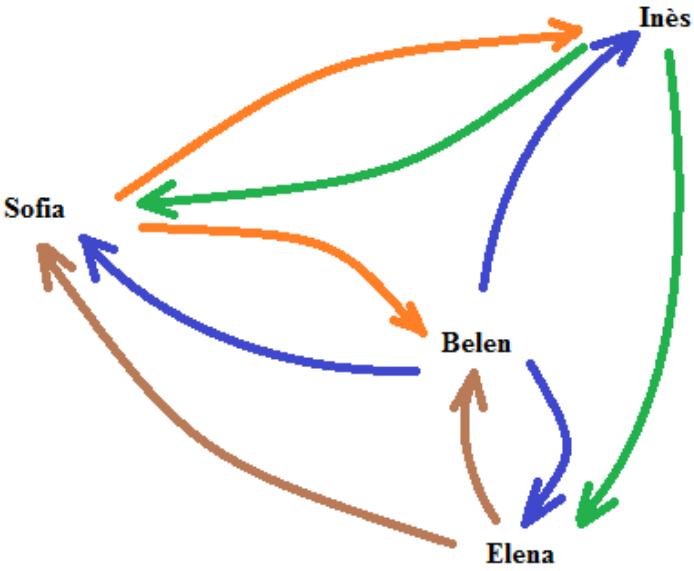
Il s'agit d'une piste complémentaire à celle des moteurs de recherche traditionnels de type Google.

La pertinence d'une page web

Jean-Alain Roddier
IA-IPR de mathématiques

L'algorithme utilisé par le logiciel Google est un algorithme complexe qui attribue à chaque page web une valeur numérique que l'on appelle sa pertinence ou PageRank. Toutes les pages web contenant ainsi le ou les mots proposés par l'utilisateur sont ensuite classées suivant cette valeur numérique et les pages renvoyées par Google sont ainsi classées par ordre décroissant de leur pertinence. C'est cette notion de pertinence qui est délicate à percevoir et que nous vous proposons de découvrir.

Prenons un exemple concret :



Considérons quatre jeunes filles dont chacune d'entre elles crée une page web identifiée par son prénom, on a ainsi 4 pages sur internet nommées : Belen, Elena, Inès et Sofia. Ces quatre copines se connaissent bien et elles ont souhaité mettre des liens sur leur page web vers la page de l'une ou plusieurs de ces copines. Pour illustrer cet ensemble de liens, on utilise ce que l'on appelle un graphe orienté (figure ci-contre). Les flèches bleues signifient ainsi que sur la page « Belen », on trouve trois liens pointant vers ses trois copines.

Si ces quatre pages étaient les seules sur la toile, on peut se demander quelle serait celle à laquelle on attribuerait une valeur plus grande que les autres.

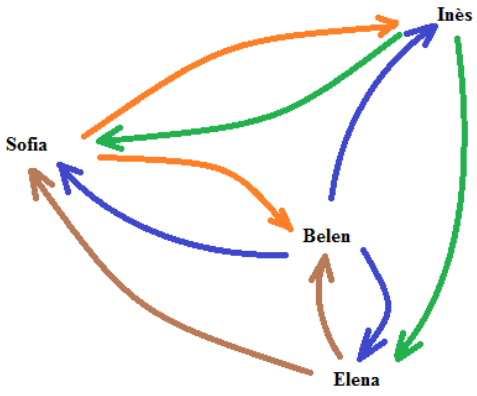
1. Les pertinences naïves

Naïvement, on peut penser que l'évaluation de la pertinence d'une page web émane d'un processus de lecture de cette page à l'image d'une professeure qui relève dans la copie d'un élève le caractère pertinent de son argumentation. Vu le nombre de pages web (environ 1000 milliards de pages), ce processus de lecture peut être qualifié de naïf. Une autre pertinence naïve pourrait consister à évaluer le nombre de pages qui pointe vers une page donnée, pourquoi est-ce naïf là-aussi ? Car tout simplement, il serait alors facile de biaiser le système en créant de nombreuses pages web vides pointant vers une page donnée afin de faire augmenter artificiellement sa pertinence.
On approche du principe de calcul du PageRank lorsque l'on essaie de contrecarrer ce biais, c'est ce que nous allons voir à présent.

2. La matrice de transition du graphe

Nous allons reprendre le graphe de nos quatre amies et lui attribuer ce que l'on appelle sa matrice de transition.

$$\begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 & 0 \end{pmatrix}$$



Expliquons un tant soit peu la construction de cette matrice, dont les 4 lignes et les 4 colonnes sont à lire dans l'ordre Belen Elena Inès Sofia ; pour ce faire observons sa deuxième ligne : **1/3 0 1/2 0** :

- le **1/3** correspond au fait que la page de Belen pointe vers Elena et que la page de Belen pointe vers trois pages. En résumé, le fait que Belen pointe vers Elena a pour poids **1/3** ;
- le **0** correspond au fait qu'Elena n'a pas créé de lien réflexif sur sa page ;
- le **1/2** est lié à la page d'Inès qui pointe vers la page d'Elena et qui pointe vers deux pages ;

- enfin le **0** exprime le fait que Sofia n'a pas créé de lien vers la page d'Elena.

3. Un utilisateur aléatoire

On considère à présent un utilisateur qui se promène aléatoirement sur la toile constituée uniquement des 4 pages de ces bonnes copines. Au départ, la probabilité qu'il soit sur une de ces 4 pages vaut ¼. On réunit ces quatre probabilités dans une nouvelle matrice X_0 qualifiée d'unicolonne présentée ci-contre.

$$\begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

Pour connaître la probabilité que notre utilisateur se trouve sur tel ou tel page après le premier parcours, il suffit de multiplier la matrice A par la matrice X_0. Nous obtenons ci-contre la matrice unicolonne X_1.

$$\begin{pmatrix} 1/4 \\ 5/24 \\ 5/24 \\ 1/3 \end{pmatrix}$$

Certaines calculatrices disposent d'un petit logiciel intégré qui permet de faire ce genre de calcul, nous pouvons alors poursuivre le processus en observant les probabilités obtenues après :

- ✓ deux parcours : $\begin{pmatrix} 13/48 \\ 3/16 \\ 1/4 \\ 7/24 \end{pmatrix}$
- ✓ trois parcours : $\begin{pmatrix} 23/96 \\ 31/144 \\ 17/72 \\ 89/288 \end{pmatrix}$
- ✓ etc

La théorie plus compliquée des matrices ergodiques permet d'affirmer que la suite des matrices unicolonnes ainsi obtenues converge. Le phénomène tend ainsi à se stabiliser vers une matrice unicolonne dont nous fournissons ici une valeur approchée obtenue à partir du calcul de X_{100}.

$$\begin{pmatrix} 0,2542 \\ 0,2034 \\ 0,2373 \\ 0,3051 \end{pmatrix}$$

Nous obtenons des valeurs approchées du PageRank de chaque page, ce qui permet de les classer dans l'ordre suivant :

Sofia, **Belen**, **Inès**, **Elena**.

Ce procédé de calcul a été mis au point par l'informaticien Lary Page (photographie ci-contre) co-fondateur de Google d'où le nom de « PageRank » pour rang d'une page obtenu en appliquant l'algorithme de Lary Page. Ceci étant, d'autres aspects complexes et plutôt secrets viennent interférer avec les résultats des calculs fournis par l'algorithme, il s'agit ainsi de faire augmenter le PageRank en intégrant des données personnelles connues sur l'utilisateur afin de lui fournir des pages répondant non seulement à ses attentes mais aussi à des objectifs bien souvent commerciaux.

La littératie numérique

Laurent Chéno
Inspecteur général de l'Éducation nationale groupe des mathématiques

Il nous est tous arrivé cette expérience fâcheuse : se retrouver privé d'internet à cause d'une connexion défectueuse, d'une panne ou d'un séjour à la campagne dans une zone non couverte par les réseaux. Et c'est alors qu'on se rend compte à quel point nous sommes accoutumés à utiliser l'internet et en particulier les moteurs de recherche dans notre vie quotidienne. Nous n'achetons plus de carte routière, ni de guide touristique : il est tellement plus simple de taper l'identification d'un lieu dans le champ de recherche de Google, qui nous donnera tout de suite la carte, le téléphone, les commentaires des visiteurs, et toute l'information utile pour notre excursion. Nous avons oublié de qui est une citation ? Nous la tapons et Google nous retrouve instantanément l'auteur. Nous ne connaissons pas le sens du mot *sérendipité* ? Google a la réponse pour nous.

Mais toute médaille a son revers : à la question *sérendipité*, Google propose 225 000 résultats. Qui ira voir les pages 2, 3, ou suivantes ? Et qui repérera dans la 27^{e} page de réponses le lien sur une page qui, justement, illustre le concept de sérendipité par la navigation sur le Web ? Dans cette abondance de réponses, chacun va finalement se limiter à la première page de réponses de Google, voire à la première réponse. De là à dire que Google nous apprend à penser, voire nous dicte quoi penser…

La littératie numérique, qui est un des objectifs de la formation au numérique que souhaite développer l'éducation nationale, désigne l'aptitude à comprendre et à utiliser les outils numériques dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités. Elle ne peut s'acquérir que grâce à une éducation aux médias et à l'information que l'École doit pouvoir offrir à chaque élève : il s'agit donc en particulier d'apprendre à utiliser un moteur de recherche, c'est-à-dire à savoir composer une requête précise, à avoir le recul critique nécessaire devant les résultats proposés, et à citer ses sources au moment de leur réutilisation.

Cet enseignement doit être l'affaire de tous : bien entendu les professeurs documentalistes sont en première ligne, mais toutes les disciplines sont sans doute directement concernées. Il suffit d'ailleurs de demander à un professeur, quelle que soit sa matière ou son niveau d'enseignement, s'il a déjà eu des copiés collés de Wikipedia sur une copie, pour comprendre que tous doivent prendre leur part dans cette éducation aux médias et à l'information.

Le document que vous tenez entre vos mains (ou que vous lisez sur un écran) est une brique utile à la construction de cette nouvelle compétence, qu'on pourrait appeler *numératie* : en déchiffrant le fonctionnement même de l'algorithme Page Rank au cœur des moteurs de recherche, il permet d'avoir une utilisation plus responsable et plus intelligente de Google et de ses homologues.

L'information, même abondante, ne suffit pas à immédiatement construire la connaissance : que les enseignants se rassurent, si leur rôle se modifie peut-être dans le cadre d'une société numérique, il reste essentiel dans l'éducation des élèves qui leur sont confiés.