Fonctionnement d'un moteur de recherche

A la lecture des différents ouvrages consacrés à la description analytique des moteurs de recherche et de leur écosystème [Simonnot, 2012 ; Mesguich et Thomas, 2013], il semble que ce que Burt Nanus décrivait en 1960 comme les quatre étapes fondamentales de tout processus de recherche d'information soit applicable au cas des moteurs de recherche, qui doivent 1) enregistrer les informations de façon à ce qu'elles puissent être mobilisées ; 2) les stocker ; 3) les sélectionner ; 4) les acheminer [Nanus, 1960, p. 279].

Les caractéristiques intrinsèques des moteurs de recherche dépendent des choix effectués par leurs concepteurs : à quoi le moteur ressemble, de quoi il est fait, les actions qu'il autorise. Pour chacune des quatre étapes citées, le concepteur doit faire des choix. Il décide de la façon dont les informations seront enregistrées, comment et où elles seront stockées, quels critères prévaudront à leur sélection, et, enfin, comment l'interface se présentera [Rieder, 2005, p. 28].

Collecter

Un moteur de recherche prend connaissance des documents en ligne sur le web. Parce que de nouvelles pages sont mises en ligne chaque seconde, il procède à cette tâche continuellement. Ses concepteurs paramètrent à cet effet des logiciels appelés « *crawlers* », « *user-agents* », « *bots* », ou encore « *spiders* » (ces différents noms désignent exactement le même type de logiciel), conçus pour visiter les documents et copier les informations qui s'y trouvent.

Les *crawlers* copient la totalité du contenu d'une page ou une partie seulement, comme par exemple les adresses URL ou les adresses courriel présentes sur la page. Leur concepteur choisit le type d'information qu'il souhaite recueillir. Dans tous les cas, les *crawlers* se déplacent grâce aux liens hypertextes reliant les pages entre elles. Une page n'étant pointée par aucun lien sera donc hors de leur portée. Tout comme les êtres humains, ils ont besoin de moyens de transport (les liens) et d'indications concernant leur destination (les « ancres », c'est-à-dire les mots sur lesquels les liens sont positionnés).

Le *crawler* doit être en mesure, en partant d'une page prédéfinie par son concepteur, de visiter les pages dont le contenu est de la meilleure qualité possible [Najork et Wiener, 2001], de visiter les pages dont le contenu est spécifique [Chakrabarti *et al*., 1999] et le format particulier [Diligenti *et al*., 2000], puis de revisiter les pages régulièrement [Cho et Garcia-Molina, 2000] et, enfin, de rendre des comptes de l'état d'avancement de son travail [Talim *et al.*, 2001].

Les éléments essentiels à la création d'un *crawler* efficace sont la flexibilité, l'optimisation coût/performance, la solidité, le respect des standards, la perfectibilité et la possibilité de reconfigurer les paramètres [Shkapenyuk et Suel, 2002, p.6]. Tout en s'assurant que son logiciel correspond à ces critères, le concepteur du *crawler* répond aux questions suivantes : Quelle est la fréquence des visites sur une même page ? Dans quel ordre les pages doivent-elles être visitées ? [Edwards *et al.*, 2001]

Chaque moteur déploie plusieurs *crawlers* simultanément, qui visitent et revisitent les documents à des fréquences variées. A chaque fois qu'ils passent sur une page, ils consomment une partie de la bande passante du serveur [Langville et Meyer, 2006, p.16-17], ce qui dans certains cas extrêmes où de nombreux *crawlers* se rendraient sur la même page au même moment peut entraîner une panne du serveur. Les *crawlers* laissent des traces numériques, dont les éditeurs ont connaissance grâce aux fichiers appelés « *logs* » ou « fichiers journaux », enregistrant par ordre chronologique les actions exécutées sur un serveur.

Protocole d'exclusion

L'éditeur d'un site web peut théoriquement empêcher les *crawlers* de visiter son contenu grâce au « protocole d'exclusion ». Initié par l'informaticien Martijn Koster en 1994 dans le cadre d'un groupe de travail de l'*Internet Engineering Taskforce*, ce protocole consiste à inscrire une ressource textuelle appelée « *robots.txt* » à la racine d'un site web. Les informations contenues dans cette ressource sont spécifiquement adressées aux *crawlers*. L'éditeur peut s'en saisir pour demander que soit exclue une partie ou la totalité de son site du champ d'indexation des moteurs.

Le *robots.txt* est un protocole volontaire destiné à formaliser des relations institutionnelles et économiques entre les éditeurs de contenus et les propriétaires des moteurs de recherche [Elmer, 2009], sans que rien ne contraigne techniquement ou juridiquement ces derniers à le respecter [Sire, 2015b]. Le concepteur qui aurait des intentions malveillantes peut d'ailleurs utiliser les informations inscrites dans le *robots.txt* pour détecter l'emplacement d'informations potentiellement sensibles [Spencer, 2009].

Sitemaps

En plus du *robots.txt*, les éditeurs peuvent utiliser le code source de leurs sites web pour adresser d'autres types d'informations aux *crawlers* et faciliter l'indexation de leurs contenus sur les moteurs de recherche. C'est le cas notamment du *Sitemap,* qui se présente sous la forme d'un sommaire, inscrit à la racine du site (au même endroit que le *robots.txt*), grâce auquel le *crawler* sait quelles sont les nouvelles pages publiées et les pages remises à jour. Il pourra ainsi se rendre directement sur ces pages depuis les liens inscrits dans le *Sitemap*, sans avoir à passer par la même arborescence qu'un internaute. Autrement dit, cette extension du *robots.txt*, proposée initialement par les ingénieurs de Google et désormais adoptée par l'ensemble des moteurs, permet d'« aplatir » la structure hypertextuelle du site pour faciliter la tâche du *crawler*.

Meta-tags

L'éditeur peut inscrire des informations dans le code source des pages destinées aux *crawlers* grâce à des *meta-tags*. Cela lui permet par exemple de demander que la page en question ne soit pas indexée et que le *crawler* ne suive pas les liens qui y figurent, ou bien de spécifier une durée d'indexation de la page, ou bien encore de préciser qu'aucune traduction de la page dans une autre langue ne devra être proposée aux internautes. Comme avec le protocole d'exclusion, ces *meta-tags* n'ont pas valeur de loi et le concepteur d'un *crawler* peut tout à fait décider de ne pas les prendre en considération.

Indexer

Après avoir pris connaissance des documents disponibles, le moteur les répertorie dans un même index en générant pour chacun d'entre eux des descripteurs : thème, date, auteur, mots-clés. Là encore, le concepteur doit faire des choix : faut-il indexer le document intégralement ou bien seulement son titre ? sa description ? les méta-informations ? [Halavais, 2009, p. 17].

« L'index forme la base sur laquelle s'exercera le traitement algorithmique qui suit, ses caractéristiques déterminent la capacité et la facilité qu'auront les algorithmes à trouver des

résultats intéressants. Cependant, la richesse de l'index est proportionnelle à son "poids" et un index qui fournit beaucoup d'informations pour chaque page qu'il référence prend non seulement plus d'espace de stockage, mais devient aussi plus lourd à traiter par la suite » [Rieder, 2006, p. 120].

Apparier requête et contenus

Dès lors que l'on peut établir automatiquement le profil des documents en ligne à la suite des procédures de *crawling* et d'indexation, la machine est en mesure de corréler les symboles ainsi extraits et les symboles fournis par la requête formulée par l'internaute (on dira que l'algorithme les *apparie*), obtenant ainsi plusieurs symboles de sortie appelés, dans leur ensemble, « résultats ». Ces résultats sont présentés sous forme de liste de liens hypertextes classés selon la pertinence supposée des documents vers lesquels ils pointent, de celui qui est potentiellement le plus pertinent à celui qui l'est le moins (tout en étant, même très peu, considéré comme étant potentiellement pertinent).

Pour que l'algorithme fonctionne, il est impératif de définir quels symboles seront considérés parmi ceux que contiennent les documents et la requête, et de les assortir de pondérations permettant de relativiser leurs importances respectives dans le calcul de pertinence. Le concepteur d'un moteur doit par conséquent faire des choix visant à ce que l'outil puisse apparier le mieux possible, selon ce qu'il considère lui-même comme étant « le mieux possible », une requête et une liste de références hypertextes.

Contenu

Les mots employés par l'internaute dans sa requête et les mots employés par l'éditeur pour rédiger son contenu sont appariés par l'algorithme. Les mots figurant dans l'adresse URL du document, dans le titre, les mots-clés associés au contenu sous forme de « *tags* », sont eux-aussi considérés par le moteur au moment de générer une liste de résultats.

Plus un mot figure tôt dans l'adresse URL, le titre d'une page ou son contenu et plus l'algorithme le considère comme un descripteur potentiellement pertinent. Les conjonctions de coordination, les déterminants et les verbes les plus usuels sont en général exclus du calcul.

Centralité

L'algorithme traite l'information enregistrée par les *crawlers* concernant les liens entrants et sortants. Il s'agit de *situer* chaque document dans la topographie hypertexte. C'est le rôle du PageRank pour le moteur *Google*, qui considère chaque lien pointant vers un document comme un indicateur de pertinence. Plus un document a été identifié comme étant pertinent, plus les liens qu'il effectue lui-même vers d'autres documents augmentent le PageRank de ces derniers.

« Le PageRank déploie une note de 1 à 10 sur une échelle logarithmique qui mesure le nombre de liens reçus par la page venant d'autres pages en considérant que les sites s'envoient les uns aux autres une force qui, dans le jargon du référencement, va très vite être appelée le "*Google Juice*" ou "jus de lien" » [Cardon, 2013, p. 76].

Le calcul du PageRank

Le PageRank « r » d'une page « A » est calculé ainsi :

$$r(A) = \frac{\alpha}{N} + (1-\alpha)\left(\frac{r(B_1)}{|B_1|} + \ldots + \frac{r(B_n)}{|B_n|}\right)$$

où $r(B_1)$, $r(B_2)$, $r(B_3)$,… $r(B_n)$: les valeurs PageRank des pages B_1, B_2,… B_n pointant vers A ;

$|B_1|$ $|B_2|$ … $|B_n|$: le nombre de lien présents sur chaque page B_1, B_2,… B_n ;

N : le nombre de pages B_1, B_2,… B_n.

et α : une constante comprise entre 0 et 1, appelée « facteur d'amortissement ».

Cette formule nous apprend que plus une page effectue de liens, moins chacun des liens est porteur de pertinence pour la page pointée. D'autre part, plus les liens entrants sont nombreux, plus la pertinence supposée de la page qui les reçoit augmente. Il y a ainsi deux effets principaux : le nombre de liens entrants et l'autorité des sources où se trouvent ces liens. Le facteur d'amortissement a lui aussi un impact significatif sur le PageRank et peut par conséquent modifier de façon substantielle le classement des documents dans le cas où il viendrait à être modifié [Langville et Meyer, 2006 ; Rieder, 2012]. Il a été fixé à 0,85 par les ingénieurs de Google.

Source

L'algorithme d'un moteur de recherche peut considérer la pertinence du site web sur lequel se situe une page en plus de considérer la pertinence de la page elle-même. Pour calculer la pertinence d'une source, différents critères peuvent être considérés comme par exemples l'historique des visites sur ce site web, le temps passé en moyenne par les visiteurs et la moyenne du PageRank de l'ensemble des pages.

Autorité

L'autorité d'un auteur, et non pas seulement d'une page ou d'une source, peut aussi être considérée. C'est ce qu'a fait Google lorsque ses ingénieurs ont ajouté à leur algorithme, en juin 2011, un critère appelé « AuthorRank », utilisé pour mesurer la popularité et la pertinence d'un auteur. Grâce à la balise « rel=author » inscrite dans le code source du document, et reliée à un compte *Google+*, ou bien grâce à un « badge Google+ », le moteur de recherche identifiait l'auteur et lui attribuait une pertinence en fonction par exemple du PageRank moyen des pages signées, du nombre de documents publiés, de l'historique des visites et du temps passé sur ses articles, de son activité sur les réseaux sociaux, et de sa présence sur *Wikipédia*, *Google Books*, *Google Scholar*.

Google a mis fin à ce dispositif de signature numérique en 2014, sans pour autant annoncer clairement que l'autorité des auteurs ne serait plus considérée dans le calcul de pertinence.

Performance

L'algorithme d'un moteur de recherche peut prendre en compte l'efficacité de l'infrastructure, et notamment la vitesse de chargement de chaque page. Le concepteur fait alors

le choix de considérer la *performance* du contenant dans le calcul de *pertinence* du contenu. C'est le cas de *Google* depuis février 2009.

Signaux sociaux

Les signaux considérés par l'algorithme d'un moteur de recherche peuvent provenir des internautes qui, lorsqu'ils interagissent avec un contenu, donnent des indications quant à sa pertinence. Les « *posts* » sur les réseaux sociaux comme *Facebook*, *Twitter* et *Google+*, ainsi que les recommandations effectuées par les internautes via des dispositifs de « *Like* », de « *+1* », ou de notation sous forme d'étoiles, peuvent être pris en compte pour le classement des documents [Kembellec et al., 2014].

Dans une interview de janvier 2012, Amit Singhal, employé de la firme Google, justifie l'attention portée aux signaux sociaux de la manière suivante : « Un bon produit ne peut être construit qu'à condition que nous comprenions qui est qui, et qui est relié à qui. Les relations sont également importantes au sein même du document. […] Car fondamentalement, il ne s'agit pas que de contenu. Il s'agit d'identité, de relations et de contenus » [Sullivan, 2012].

Interventions manuelles

Les signaux d'entrée de l'algorithme peuvent venir de personnes chargées de juger de la pertinence des pages web. Il a ainsi été dévoilé en 2011 que des travailleurs à domicile, employés par les entreprises Leapforce, Butler Hill et Lionbridge, jouaient un rôle manuel dans le processus de sélection et de hiérarchisation opéré par *Google*. Les employés de ces entreprises visitent certains sites web dont les coordonnées leur sont communiquées par Google et attribuent une note aux documents dont ils prennent connaissance. Pour chaque page, il leur est demandé de déterminer si elle est : « *vital* », « *useful* », « *relevant* », « *slightly Relevant* », « *off-topic or useless* » ou « *unratable* ». Pour la source, les travailleurs à domicile jugent de la réputation du site : bonne, mauvaise, ambiguë, OK, malicieuse ou impossible à déterminer.

Autres critères

Le nombre de critères aujourd'hui pris en compte par *Google* pour apparier une requête et des contenus est compris entre 200 et 300 [Richard, 2011], parmi lesquels, outre les critères nommés précédemment, on compte l'historique du taux de clics, le temps passé, la fraîcheur des documents, la fréquence de publication et la fréquence des mises à jour. Même si les professionnels du secteur peuvent avoir une idée des facteurs déterminants du classement des documents, nul ne connaît les critères ou leurs pondérations dans le détail.

Les changements opérés régulièrement par les ingénieurs de Google, qui définissent de nouveaux critères et réévaluent les pondérations, rendent tout projet d'étude empirique du comportement du moteur extrêmement difficile, voire impossible.

Personnalisation

L'algorithme d'un moteur de recherche peut se comporter différemment selon ce que le moteur « sait » de l'utilisateur [Kembellec *et al.*, 2014]. L'appariement est effectué alors en fonction de critères liés non plus seulement au contenu, au contenant, aux liens hypertextes,

aux signaux sociaux, à l'auteur ou à l'éditeur, mais également à l'individu en quête d'information. Les moteurs comme *Google* essayent ainsi de calculer la pertinence de tel ou tel document aux yeux de tel ou tel individu, de façon à procéder à des hiérarchisations personnalisées [Goldman, 2010].

L'interface

Une fois qu'il a effectué l'ensemble des choix concernant le *crawling* et l'indexation, puis paramétré l'algorithme, le concepteur d'un moteur doit faire de nouveaux choix concernant cette fois l'apparence de l'outil qu'il mettra à la disposition des internautes. Les moteurs présentent en général une interface simple où une barre permet de formuler une requête. Depuis sa création, Google a gardé un *design* épuré : fond blanc, logo, fonction « Recherche », et une fonction « J'ai de la chance », conduisant l'utilisateur directement sur la page identifiée comme étant la plus pertinente, sans passer par la liste de résultats.

Suggestions

Lorsque l'internaute est en train de formuler sa requête, dès qu'il a tapé les premières lettres sur son clavier, le moteur peut lui suggérer plusieurs requêtes déjà formulées par d'autres internautes. L'internaute peut alors continuer de taper sa propre requête, ou choisir parmi les requêtes suggérées.

Google va plus loin, en proposant des résultats sous la barre de recherche pendant que l'internaute formule sa requête. Ce dispositif, nommé *Google Instant*, permet de proposer des résultats différents à chaque fois que l'internaute ajoute une lettre ou un signe de ponctuation à sa requête, jusqu'à ce que, finalement, une liste se stabilise quand l'internaute a fini de frapper les touches de son clavier.

Liens simples et enrichis

Les liens générés par un moteur apparaissent en bleu et sont accompagnés d'une description, en noir, et de l'adresse URL de la page pointée, en vert. Un clic sur le lien conduit directement vers la page pointée.

Différentes fonctionnalités accompagnent le lien, comme le « cache », qui permet d'accéder non pas au document enregistré sur les serveurs de l'éditeur, mais au document enregistré par les *crawlers* sur les serveurs du moteur de recherche. En utilisant cette fonction, l'internaute ne sera pas comptabilisé parmi les visiteurs du site. Et dans le cas où un éditeur a supprimé ou modifié un document enregistré par le moteur, l'internaute accédera quand même au document (s'il s'agit d'une suppression), ou à sa version non modifiée (dans le cas d'une modification). Il est toutefois possible pour un éditeur de spécifier aux *crawlers*, grâce au *robots.txt* ou au *meta-tag* spécifique « *noarchive* », qu'il ne souhaite pas qu'une telle fonctionnalité soit proposée.

D'autres fonctionnalités peuvent être proposées, comme par exemple la fonction « Pages similaires » sur *Google*, qui permet à l'internaute de relancer la recherche en ayant spécifié que c'est ce type de contenu en particulier qui l'intéresse (e-commerce, presse en ligne, etc.).

Les liens peuvent également être « enrichis » et comporter par exemple une photo, une indication des notations attribuées par les internautes, une liste de sous-liens, un ordre de prix

pour les restaurants, la durée de préparation pour une recette, la liste des titres d'un album de musique, la date d'un événement, etc.

Liens sponsorisés

Des « liens commerciaux » ou « liens sponsorisés » peuvent apparaître sous la même forme que les liens décrits précédemment, mais dans des encarts spécifiques en haut et à droite de la page de résultats, et parfois, plus rarement, en bas. Un clic sur un de ces liens déclenche un paiement de l'éditeur vers le site duquel l'internaute est dirigé au propriétaire du moteur sur lequel le lien a été généré. Les liens non sponsorisés sont quant à eux appelés « liens organiques » ou « liens naturels ».

Moteurs verticaux

Il existe des moteurs de recherche spécialisés, appelés aussi « moteurs verticaux », dont le champ d'indexation n'est consacré qu'à un seul format de document (images, vidéos, feuilles de calcul, PDF, etc.), ou à un champ d'indexation restreint (par exemple les sites web identifiés au préalable comme étant consacrés au traitement de l'information d'actualité), ou à un type d'information en particulier (prix des produits vendus sur les sites de e-commerce, horaires des séances de cinéma, etc.).

Ces moteurs verticaux peuvent être intégrés à un moteur généraliste, qui, dans le cas où le dispositif aura jugé qu'une telle action était supposément pertinente, proposera à l'utilisateur de spécifier sa requête : en cliquant sur un lien hypertexte apparu dans les résultats, l'internaute signalera que seules les images l'intéressent et sera redirigé vers le moteur vertical concerné. Il est également possible de faire apparaître les principaux résultats des moteurs verticaux dans la liste du moteur généraliste. Ce dernier est ainsi capable de répondre à des questions extrêmement spécifiques en se basant sur des procédures algorithmiques conçues expressément pour tel ou tel type de requête et effectuées au sein de champs d'indexation circonscrits *a priori*.

Google et la recherche universelle

Le moteur Google Search ne peut pas traiter toutes les requêtes pareillement. Les ingénieurs s'en sont aperçus le 11 septembre 2001. Alors que l'effondrement des Tours jumelles était relayé en direct par les chaînes de télévision du monde entier, et que les stations radio avaient interrompu leurs programmes pour commenter la catastrophe, les internautes effectuant sur Google la requête « *world trade center* » étaient redirigés vers des pages qui ne mentionnaient pas les attentats en cours [Wiggins, 2001].

Dans les mois qui suivirent, un employé de Google, Krishna Bharat, décida de spécifier un champ d'indexation où ne seraient intégrés que des sites identifiés au préalable comme étant spécialisés dans le traitement de l'information d'actualité [Vise, 2005]. Sur ces sites, les *crawlers* augmenteraient la fréquence de passage de sorte que le moteur puisse indexer les articles quelques minutes après leur publication. C'est ainsi que Google Actualités vit le jour [Sire, 2015a]. Dès lors qu'une requête serait effectuée par un internaute sur le moteur de recherche généraliste, et que le dispositif identifierait qu'elle était possiblement liée à un événement d'actualité, un transfert serait proposé depuis la liste de résultats produite par le moteur généraliste Google Search vers une liste de résultats produite par le moteur spécifique Google Actualités. Le fait de proposer à l'internaute de transférer sa requête vers un type de

contenu spécifique permettait de spécifier la nature de son besoin d'information et de maximiser les chances de le satisfaire.

Google appliqua la même recette à d'autres types de contenus. Grâce à une stratégie combinant développement interne et croissance externe, Google mit notamment en place les moteurs spécialisés « Images », «Vidéos », « Scholar » (contenus académiques), « Maps » (localisations et les itinéraires), « Books » (livres), « Shopping » (e-commerce), « Flight » (horaires de vols), « Movies » (horaires de cinéma). Chacun de ces outils s'est vu attribuer une adresse URL permettant aux internautes de se rendre directement sur l'un d'entre eux pour effectuer une requête.

D'autre part, les ingénieurs de Google paramétrèrent le moteur généraliste de façon à ce qu'il interroge systématiquement les moteurs spécifiques et qu'il puisse suggérer à l'internaute un transfert à chaque fois que cela lui semblerait pertinent. En outre, ils firent en sorte que le moteur généraliste puisse afficher directement dans ses résultats les principales informations remontées par les moteurs spécifiques. Dans le cas où ces informations suffiraient à répondre à la requête de l'internaute, celui-ci n'aurait donc même pas à effectuer un transfert vers le moteur spécifique. C'est ce qu'on appelle la « recherche universelle » : Google Search ambitionne de pouvoir répondre à tout type de requête, même extrêmement particulier. Il est de moins en moins généraliste et de plus en plus complet, attentif à des spécificités qu'un moteur unique, ne communicant pas avec des sous-moteurs, ne serait pas en mesure de traiter.

Paramètres de recherche avancée

Le concepteur d'un moteur de recherche peut laisser à l'internaute la possibilité d'accéder à des paramètres de recherche avancés, visant par exemple à spécifier une plage temporelle ou un lieu, inclure un opérateur booléen, n'afficher les résultats que d'un seul site, spécifier un type de fichier, une langue, etc. Des interfaces sont créées pour que l'internaute puisse effectuer ce type de recherche (exemple : google.fr/advanced_search).

Par ailleurs, certains signes de ponctuation, symboles ou mots renvoient à des opérateurs spécifiques et permettent à l'utilisateur d'effectuer une recherche avancée directement depuis la barre du moteur. On peut ainsi exclure des mots, lier des mots, chercher des pages effectuant des liens vers un site en particulier, etc., sans passer par l'interface de recherche avancée. Par exemple, l'utilisation de guillemets dans la barre du moteur permet de spécifier qu'on cherche une expression en entier, comme s'il s'agissait d'un seul mot : « "la couleur du cheval d'Henri IV" » permettra de faire remonter dans les résultats les pages qui contiennent ces mots, ensemble, dans cet ordre. Autre exemple, la requête « inurl:arthur inurl:cheval filetype:henri » permettra de retrouver les documents au format PDF dont l'adresse web contient les mots *cheval* et *henri*. Troisième exemple, la requête « Charles Péguy site:www.lemonde.fr » permettra de trouver les pages du site web du journal *Le Monde* où l'auteur Charles Péguy est cité.

Les principaux opérateurs interprétés par Google

Opérateur	Utilité	[Exemple de requête]
« »	rechercher une expression exacte.	["La guerre de cent ans"] fait remonter les sites où les mots *La guerre de cent ans* sont présents dans cet ordre.

-	exclure un terme.	[Smartphone -iPhone] fait remonter les pages concernant les Smartphones où l'iPhone n'est pas mentionné.
OR	rechercher une page où se trouve un terme ou l'autre.	[Google OR Yahoo] fait remonter les pages contenant l'un ou l'autre de ces deux termes.
site:	rechercher les pages web d'un site spécifique.	[Printemps arabe site:lemonde.fr] ne fera remonter que les pages hébergées par le site du journal *Le Monde* évoquant le Printemps arabe.
link:	rechercher toutes les pages qui redirigent vers une page en particulier.	[Printemps arabe link:lemonde.fr] fait remonter les pages évoquant le Printemps arabe et pointant vers lemonde.fr.
filetype:	limiter la recherche au type de fichier spécifié	[Guerre de cent ans filetype:pdf] ne fera remonter que les documents PDF consacrés à La guerre de cent ans.
define:	obtenir la définition d'un terme	[define:contingence] donnera la définition du mot « contingence ».
weather:	obtenir la météo d'un lieu.	[weather:toulouse] donnera la météo pour la ville de Toulouse.

Les caractéristiques informationnelles et communicationnelles de ce qui est produit

Un moteur produit pour chaque requête une liste de documents supposés pertinents et un classement de ces documents. Dès lors, il ne peut pas être cantonné au statut d'objet technique ; il est aussi, et autant, une technologie politique [Introna et Nissenbaum, 2000] positionnée à un niveau méta-éditorial [Chartron, Rebillard, 2007]. Par son action, certains éditeurs, certains auteurs, certains sujets et certains traitement éditoriaux deviennent « plus publics » que les autres [Cardon, 2010]. Dès lors, il convient de « poser aux moteurs de recherche comme Google la même question que celle que les chercheurs ont posé aux médias traditionnels : les voix sous-représentées et les points de vue différents peuvent-ils être entendus étant donné le filtre des moteurs de recherche ? Quel rôle la publicité joue-t-elle dans les résultats ? Un petit nombre d'acteurs dominent-ils l'industrie ? A la seule condition de répondre à ces questions, nous pourrons juger de la réelle capacité « délibérative » du Web » [Diaz, 2008, p. 15].

Interaction sociotechnique

La seule action du concepteur ne suffit pas à expliquer ce qui apparaîtra à l'écran d'un internaute ayant formulé une requête sur un moteur. La liste de résultats dépend en effet également des actions des éditeurs et des internautes. Elle dépend des mots choisis par les internautes. Elle dépend des choix effectués par les éditeurs. Finalement, le moteur produit un lien *entre* ces acteurs et leurs actions : c'est une interaction sociotechnique. C'est pourquoi il serait faux de dire que le procédé est purement technique. Comme le souligne Engin Bozdag : « les services de *gatekeeping* en ligne ne sont pas constitués seulement de machines utilisant des algorithmes ; ils résultent d'un mélange entre les actions d'éditeurs humains et les actions des lignes de code informatique écrites par des humains » [Bozdag, 2013, p. 224].

Parcours de navigation

La liste de résultats produite par un moteur n'est pas seulement un ensemble de liens pointant vers des pages. Se limiter à une telle assertion reviendrait à considérer que la navigation ne dépasse jamais le niveau « n+1 », c'est-à-dire qu'elle s'arrête aussitôt que l'internaute a cliqué sur un lien généré par le moteur. Il n'y a pas seulement des pages derrière la liste de liens, mais aussi, sur ces pages, de nouveaux liens, pointant eux-mêmes vers de nouvelles pages, contenant de nouveaux liens, etc.

Lorsqu'on considère l'arborescence du web, chaque lien constitue un choix conduisant à d'autres choix possibles. Générer une liste de liens revient donc à produire un ensemble de possibilités d'action ouvertes elles-mêmes sur d'autres possibilités d'action. Lorsqu'un internaute clique sur un lien généré par un moteur, il choisit un parcours qui le conduit à d'autres parcours possibles ; il quitte pour intégrer ; il ferme pour ouvrir. Chacun des parcours correspond à différentes opportunités communicationnelles et économiques pour le propriétaire du moteur, l'internaute et les éditeurs [Sire et Rieder, 2015].

Les caractéristiques économiques de ce qui est produit

Du point de vue de l'internaute : bien public

Le nombre de requêtes que l'on peut effectuer sur un moteur est illimité. Autrement dit, générer une liste de résultats ne diminue pas la quantité de listes de résultats disponibles. En économie, on dit que le service rendu à l'internaute a les caractéristiques d'un bien *non-rival* et que le coût marginal pour satisfaire un internaute supplémentaire est nul.

Tous ceux qui disposent d'une connexion Internet peuvent effectuer une requête sur un moteur. Le service consommé est donc également *non-exclusif*. Ces principes de non-rivalité et de non-exclusion confèrent au moteur le statut de *bien public pur* [Sonnac, 2009]. Cela n'est vrai toutefois que du point de vue de l'utilisateur.

Du point de vue de l'éditeur : positions privatives

Du point de vue de l'éditeur d'un site web, le nombre de positions disponibles dans les listes de résultats d'un moteur de recherche est potentiellement illimité. Cependant, si nous ne considérons que la première page de résultats, qui reçoit plus de 90% des clics [Optify, 2011], le nombre de positions y est bel et bien limité. Occuper une position sur cette page, pour un éditeur, aboutit à une diminution du nombre de positions disponibles pour les autres. Ainsi, de leur point de vue, on peut considérer que le service fourni par le moteur a les caractéristiques d'un bien *rival*.

Le propriétaire d'un moteur peut éventuellement faire payer les éditeurs pour apparaître dans les résultats. Ce n'est pas parce que les principaux acteurs présents sur le marché ne le font pas que cela n'est pas possible ; il n'est pas inimaginable en effet que l'intégration d'un site web au champ d'indexation de Google soit un jour facturé à l'éditeur, comme l'a déjà fait Google en faisant payer les commerçants dont les produits sont référencés sur *Google Shopping*. D'autre part, le propriétaire d'un moteur de recherche peut tout à fait exclure certains sites du périmètre d'action des *crawlers*. Considérées sous cet angle, les positions dans la liste de résultats ont donc également les propriétés de biens *exclusifs*.

Dès lors qu'il y a rivalité et exclusivité, la liste de résultats produite par un moteur de recherche peut être considérée comme un ensemble de positions *privatives* du point de vue des éditeurs.

Du point de vue des internautes et des éditeurs : liste expérientielle

Un bien d'expérience, ou *bien expérientiel*, en économie, est un bien dont on ne connaît la valeur qu'après l'avoir consommé. Du point de vue des internautes et des éditeurs, la liste de résultats produite par un moteur de recherche a les caractéristiques d'un tel bien. En effet, tant qu'il n'y a pas eu de clic de la part d'un internaute, aucun des acteurs impliqués ne peut savoir s'il bénéficiera ou non de cette liste en particulier. Il existe donc une incertitude quant à la qualité du service par un moteur de recherche jusqu'à ce que la liste ait effectivement été consommée, c'est-à-dire jusqu'à ce qu'elle ait disparu de l'écran de l'internaute.