LE POUVOIR NORMATIF DES MOTEURS DE RECHERCHE SUR LES CONTENUS

Les éditeurs de sites web n'ignorent pas que les moteurs existent et que l'un d'entre eux jouit d'une position hégémonique. Ils n'ignorent pas que sur le web « exister, c'est être indexé par un moteur » [Introna et Nissenbaum, 2000, p. 171]. Et ils n'ignorent pas que la visibilité de leurs documents dépend de leurs actions : sujets traités, structure du site, vitesse de chargement, formats, mots, liens, code [Andrieu, 2013]. Certains éditeurs agissent par conséquent en fonction de ce qu'ils savent, ou croient savoir, à propos de l'algorithme de Google. Ils *optimisent* leurs contenus dans le but de maximiser leurs chances de recevoir un trafic substantiel en provenance du moteur [Bar-Illan, 2007].

Grâce aux différentes techniques regroupées sous le terme « *Search Engine Optimization* » (SEO), « il est possible pour un individu isolé de réussir à positionner un site web tout en haut des listes de résultats de manière durable et sans rien faire d'illégal » [Boutet *et al*., 2012, p.457]. Ces techniques ont donné lieu à un secteur activité où des professionnels vendent aux éditeurs des prestations dont l'objectif est d'augmenter la visibilité de leurs documents dans les résultats de *Google* [Domenget et Michel, 2015]. Certains éditeurs décident d'internaliser en embauchant directement un spécialiste du référencement qui travaille au quotidien avec les équipes rédactionnelles, commerciales et techniques [Sire, 2014].

Le concepteur d'un moteur utilisé massivement par les internautes ne peut agir en pensant que les éditeurs ne calibreront pas leurs actions en fonction de ce qu'ils savent des siennes. Ce que le concepteur fait aujourd'hui influencera nécessairement ce que les éditeurs, eux-mêmes, feront demain, et donc la nature des documents accessibles en ligne et des liens hypertexte permettant aux internautes de circuler parmi ces documents. Autrement dit, les actions des concepteurs d'un moteur comme *Google* sont susceptibles d'influencer la nature du web tout entier. Il faut imaginer un projectionniste qui, selon sa façon de braquer la lumière sur tel ou tel point de la scène, et selon ce que les comédiens savent au sujet de *ce qui compte* à ses yeux, influencerait les déplacement et les répliques des comédiens, si bien que la pièce elle-même serait influencée par son *pouvoir*, dont dépendrait (pas totalement, mais en partie) ce que les spectateurs voient, entendent, ressentent et comprennent.

Si les algorithmes des moteurs de recherche étaient parfaitement transparents, les éditeurs seraient nombreux à agir exactement de la même façon pour optimiser le calcul de pertinence de leurs documents, auquel cas l'algorithme ne pourrait plus fonctionner, incapable de choisir parmi des éditeurs qui auraient conçu pareillement des documents traitant d'un même thème. Il faudrait trouver d'autres critères pour les départager, et ne pas divulguer ces critères. C'est la principale raison pour laquelle Google ne dévoile pas les détails de son algorithme. En revanche, la firme publie des recommandations sur son « Centre d'aide aux webmasters » destinées aux éditeurs qui souhaiteraient maximiser leurs chances de figurer en bonne place dans les résultats. Pour autant, Google ne promet rien à celui qui obéirait à la lettre à ses recommandations. L'hypothèse d'un trafic substantiel est une façon pour les ingénieurs de Google de *faire faire* aux éditeurs de sites web certaines actions que personne d'autre ne pourrait faire à leur place et grâce auxquelles le moteur peut fonctionner au plus près de ce que ses concepteurs souhaitent : les *crawlers* se déplacent plus vite, consomment moins de bande passante et identifient rapidement les pages pertinentes.

De nombreux éditeurs se conforment à la norme de publication ainsi promue par Google. Certains auteurs n'hésitent pas à comparer cette situation à un « régime disciplinaire » [Röhle, 2009] censé bénéficier aux utilisateurs du moteur, certes, mais qui va également dans le sens des intérêts économiques de Google, dès lors que ce *faire faire* permet de fluidifier l'indexation des documents en diminuant les coûts en bande passante. Il serait faux pour autant de considérer qu'est exercée ici un pouvoir totalisant. Google ne peut pas forcer les éditeurs à quoi que ce soit. Certains acteurs détournent les règles et piègent le moteur. Les ingénieurs sont

alors obligés de revoir les critères et les pondérations de leur algorithme, s'ils veulent que le dispositif continue de générer des classements pertinents.

L'information optimisée

Le sujet

Les sujets n'ont pas tous les mêmes probabilités de succès auprès des utilisateurs des moteurs de recherche. Créer un contenu à propos des sujets les plus recherchés (pornographie, météo, *people*, orthographe) maximise les chances de recevoir un trafic substantiel. Il existe des outils grâce auxquels les éditeurs peuvent connaître les sujets qui intéressent les utilisateurs des moteurs et auxquels ils ont intérêt à consacrer des pages. C'est notamment le cas de « *Tendances de Recherche* », un instrument mis à disposition des éditeurs gratuitement par Google dans le but de leur indiquer quels sont les sujets les plus recherchés sur le moteur le plus fréquenté.

Les moteurs peuvent ainsi se faire l'écho de la demande de leurs utilisateurs, et conduire les éditeurs à effectuer leurs choix rédactionnels en fonction de cette demande. Un éditeur qui suivrait entièrement cette logique abandonnerait la prérogative consistant à décider des sujets à traiter, et plus généralement à définir la ligne éditoriale de son site, pour être en permanence en train de répondre aux requêtes effectuées par les utilisateurs des moteurs de recherche.

Les mots

Pour permettre aux moteurs d'identifier facilement de quoi il est question sur une page web, le contenu doit être explicite, informatif plutôt qu'incitatif, et des mots-clés désignant sans ambigüité le thème traité doivent figurer le plus tôt possible dans le contenu textuel [Richard, 2011]. C'est pourquoi on voit fleurir sur le web des titres ayant la forme « DSK : l'affaire du Carlton », « Ukraine : les attaques ont repris », « *Mot-clé* :… ». Les moteurs de recherche sont en effet paramétrés de façon à considérer qu'il existe une forte probabilité pour que le mot situé avant les deux points d'un titre soit un descripteur pertinent de la thématique abordée. D'autre part, ces mots-clés doivent être idéalement les mêmes que les mots utilisés par les internautes lorsqu'ils formulent des requêtes auxquelles le document peut constituer une réponse pertinente.

L'adresse URL et les *tags* donnent également des informations au moteur concernant la thématique d'un document. Les exigences liées au référencement peuvent donc influencer aussi bien le contenu que les méta-contenus : pour une stratégie d'optimisation idéale, le mot-clé employé par l'internaute dans sa requête devra se trouver *le plus tôt possible* et *à la fois* dans l'adresse URL, les *tags*, le titre, le sous-titre, le chapeau et le corps d'un texte supposé pouvoir répondre à cette requête.

Quand Google conduit les journalistes à la faute

Dans un article du *Poynter*, la journaliste Kelly McBride a révélé quelles pouvaient être les dérives de l'influence du moteur Google sur le choix des mots [MacBride, 2010]. Elle a pris pour cela l'exemple du traitement journalistique d'un projet de construction d'un centre musulman à deux pâtés de maison et demi du *Ground Zero* à New York, erroné à cause l'usage de « *Tendances de Recherche* ». Puisque les internautes formulaient la requête « *ground zero*

mosque » sur Google, de nombreux journalistes utilisaient ces mêmes termes dans les titres de leurs articles à des fins de captation du trafic ; seulement voilà : l'information donnée par les titres sous-entendait qu'une mosquée serait effectivement construite sur le *Ground Zero*, ce qui était faux puisque la mosquée devait être construite plus loin (les deux pâtés de maison étant de tailles considérables) et que le projet, porté par des musulmans modérés, n'était pas seulement de construire une mosquée, mais également un centre ouvert à tous comprenant une piscine, un terrain de basket, une salle de gym, un auditorium de 500 places, une librairie, un studio d'art, un restaurant et une école de cuisine.

Cette anecdote permet de pointer la limite de la stratégie qui consiste à reprendre dans le titre d'une page web les termes utilisés par les internautes lors de leurs requêtes sur les moteurs de recherche — stratégie pourtant plébiscitée par certains journalistes [Richmond, 2008 ; Niles, 2010]. En effet, l'objectif de la requête est d'exprimer un besoin d'information, tandis que le but d'un article est de traiter l'information. La première est une question, le deuxième une réponse. Dès lors, la méthode qui consiste à reprendre dans la réponse les termes utilisés dans la question, parfaitement valable du point de vue du *Search Engine Optimization*, peut conduire à des erreurs inacceptables du point de vue journalistique.

Le contenant

Google prend en compte l'efficacité de l'infrastructure dans le calcul de pertinence, et notamment la vitesse de chargement de chaque page depuis février 2009. La firme justifie la prise en compte de ce critère en expliquant qu'elle a toujours elle-même été « obsédée par la vitesse » (sic.). Ainsi, elle fait de la *performance du contenant* un critère de *pertinence du contenu*. On ne considère pas que l'internaute est à la recherche d'une information qui ne serait « que » pertinente, mais aussi d'une information qu'il pourra consulter *confortablement* et *rapidement*. C'est un peu comme si on demandait à deux philosophes supposément compétents pour parler de Spinoza de courir un cent mètres, et que l'on décrétait que le plus rapide est sans doute le plus compétent.

Les éditeurs désireux de réduire la vitesse de chargement de leurs sites doivent investir dans des infrastructures et créer des documents dans un format qui ne sera pas trop gourmand en bande passante. Cela représente des coûts et a donc tendance à avantager ceux qui ont les moyens d'investir, comme par exemple les éditeurs capables de payer les services d'un « *Content Delivery Network* » (CDN) afin que leurs informations soient servies à l'internaute, grâce à un mécanisme de routage, par le nœud le plus proche dans le réseau Internet. Pour le coût d'une telle opération, les prix pratiqués par une entreprise comme *OVH* (spécialiste de solutions de CDN) varient de 9,99 € à 599,99 € euros par mois.

Code source

Il existe de nombreuses façons pour un éditeur de s'adresser au moteur de recherche depuis le code source de son site web, aussi bien pour maximiser ses chances d'apparaître dans les listes de résultats que pour contrôler l'apparence du lien qui finalement apparaîtra dans ces listes. Ainsi, même si ce n'est pas toujours visible pour l'internaute, l'existence des moteurs a une influence indéniable sur le code.

Le *robots.txt* et les *meta-tags* permettent aux éditeurs de contrôler l'action des *crawlers*. Les données structurées (microdonnées, microformats, RDFa), quant à elles, permettent d'enrichir les liens qui apparaissent sur le moteur. Par exemple, pour un contenu ayant fait l'objet d'une notation par les internautes, l'utilisation des données structurées permet de faire apparaître cinq étoiles dans les résultats du moteur à côté du lien pointant vers le contenu

concerné, et d'indiquer grâce à ces étoiles la moyenne des notes reçues par le contenu en question (sur cinq).

Formats

Les moteurs de recherche sont conçus principalement pour des documents rédigés en format textuel HTML. Les autres formats sont moins appropriés au travail des *crawlers*, ce qui conduit les éditeurs souhaitant être visibles dans les listes de résultats à privilégier systématiquement le langage HTML.

Pour les images, les formats les plus courants sont JPEG, PNG et GIF. Ils sont reconnus par les moteurs mais doivent idéalement être accompagnés de contenu textuel. Cependant, Google a mis au point en 2010 un nouveau format pour les images, appelé « WebP », plus léger et, donc, plus facile à indexer pour son moteur. En inventant et en publiant ce format, Google a promu une norme de publication, et a invité plus ou moins explicitement les éditeurs désireux de maximiser leurs chances d'êtres visibles sur son moteur à adopter ce format au plus vite.

Arborescence

On désigne par le terme « arborescence » l'ensemble des parcours de navigation qu'il est possible d'effectuer depuis la page d'accueil d'un site web pour atteindre les autres pages. Une arborescence trop profonde gêne les *crawlers* des moteurs de recherche, qui se déplacent à partir de la page d'accueil dans un temps limité, et ont besoin d'accéder le plus rapidement possible aux pages les plus pertinentes. Il est donc recommandé de privilégier une arborescence peu profonde.

En outre, pour indiquer aux *crawlers* l'emplacement des pages les plus pertinentes dès leur arrivée sur le site, les éditeurs peuvent avoir recours à un *Sitemap*. Il s'agit d'une extension du *robots.txt* proposée initialement par Google et désormais adoptée par l'ensemble des moteurs. Le *Sitemap* vient « en complément du *crawl* habituel ; il n'est donc pas indispensable d'utiliser ce système pour être indexé […], par contre il peut éventuellement aider à mieux indexer les sites » [Duffez, 2005]. Le *Sitemap* se présente sous forme de sommaire de liens hypertextes permettant au *crawler* de savoir sur quelles pages il devra se rendre en priorité.

Maillage

Contrairement aux autres critères de l'algorithme, le PageRank est connu dans le détail. Les éditeurs peuvent par conséquent ajuster leurs actions étant donné ce qu'ils savent de l'analyse topographique effectuée par Google. Ils savent notamment que les liens entrants apportent de l'autorité et que les liens sortants en redistribuent. Plus les liens entrants sont nombreux, plus la pertinence supposée d'un document augmente. Plus les liens sortants sont nombreux, moins l'autorité apportée par chacun de ces liens est grande. Donc les éditeurs ont intérêt à pointer vers leurs propres pages et à recevoir des liens provenant de pages extérieures.

Le PageRank encourage un certain égocentrisme hypertextuel, qui constitue pourtant un contre-sens du point de vue de Google, car si tous les éditeurs n'effectuaient des liens que vers eux-mêmes, cela créerait des bulles hypertextuelles étanches, ce qui non seulement affaiblirait la pertinence de la mesure effectuée par PageRank, mais ce qui, en outre, empêcherait les *crawlers* de trouver les documents, dès lors qu'ils se déplacent de lien en lien pour trouver de nouveaux documents. Ainsi, ce que suggère de faire la formule mise au point par les ingénieurs de Google peut nuire au fonctionnement du système dont pourtant cette formule est l'épicentre.

Si aucun éditeur n'en connaissait l'existence, l'algorithme PageRank pourrait fonctionner sans que cela n'ait d'influence sur le nombre de liens. Mais dès lors que les éditeurs

connaissent la formule, et que Google est en position dominante sur le marché des moteurs de recherche, ils sont incités à ne pas faire de liens vers l'extérieur alors que cela va contre l'intérêt de ceux qui ont mis au point la formule. C'est pourquoi Google a cherché à réduire l'influence du PageRank sur l'usage des liens hypertextes en annonçant notamment que les liens sortants pertinents augmenteraient la pertinence supposée, et en annonçant que l'autorité redistribuée par un éditeur à soi-même serait moins forte que l'autorité distribuée par un site extérieur.

Ancres

Le texte sur lequel les liens hypertextes figurent est un indicateur de la pertinence, non pas du document sur lequel le lien se trouve, mais du document vers lequel le document pointe. Ce texte est appelé « ancre ». Il est préférable du point de vue du moteur que les ancres décrivent le mieux possible le contenu pointé, plutôt que d'éditer des ancres du style : « *cliquez ici* ». Idéalement, les ancres contiennent les mêmes mots-clés que ceux qui sont employés par les utilisateurs pour effectuer une requête sur un moteur de recherche, dès lors que le document pointé par le lien fixé sur l'ancre pourrait constituer une réponse pertinente à cette requête en particulier.

Interactions sociales

Le nombre de « *Like* » Facebook, de « *Tweets* », de « +1 » ou le nombre de commentaires sont des données publiques, directement accessibles sur la page web. Les moteurs peuvent par conséquent utiliser ces données comme autant d'indicateurs de *succès* d'une page web, et, donc, de *pertinence*. Les éditeurs sont ainsi incités à encourager les interactions sociales que leur contenu est susceptible de provoquer, non pas forcément (ou pas seulement) parce qu'ils souhaitent que les internautes interagissent, mais parce que de telles interactions sont susceptibles d'augmenter la visibilité de leurs documents sur les moteurs de recherche.

Chapeaux blancs et chapeaux noirs

Les éditeurs de contenus qui se conforment strictement aux préconisations de Google sont appelés « *white hats* » (chapeaux blancs). Ceux qui au contraire contournent volontairement les règles dans le but de faire remonter des contenus qui, selon les ingénieurs de Google, ne sont pas les plus pertinents, sont appelés « *black hats* » (chapeaux noirs). Il semble ainsi que le pouvoir normatif des concepteurs de moteurs de recherche, et de Google en particulier, passe par une forme de code de conduite adressé aux éditeurs de contenus, et que celui-ci puisse être respecté ou contesté.

Dans les faits, la limite entre *black* et *white hats* est floue. Une ambigüité pèse notamment sur les pratiques que le terme *black hat* est censé désigner : « Pour la société Google et les référenceurs, ce terme recouvre l'ensemble des procédés qui biaisent les algorithmes de moteurs de recherche. Pour les autres, il rejoint celui de nuisance, de *spamming*, de messages et commentaires non sollicités sur leurs supports, […]. Pratiquement, on remarque que le terme englobe un large panel de pratiques qui pour la plupart visent au profit économique. Ainsi, sur les forums dédiés au *black hat SEO*, se mêlent les personnes souhaitant favoriser l'indexation de leur site internet par tous moyens, mais aussi des individus cherchant le bénéfice par les manières les plus illégales qui soient, tels que le piratage de sites internet, de cartes de crédit, aussi, il n'est pas étonnant que cette activité ait mauvaise presse » [Boutet et Amor, 2010, p. 195].

Google Bombing

Les « bombes Google » désignent une manière de manipuler le moteur pour faire passer un message. Il s'agit de créer de nombreux liens hypertextes visant à faire systématiquement remonter un contenu dans les listes de résultats. Pour cela, l'action doit être collective : les internautes se coordonnent pour éditer des liens vers une même page en faisant en sorte que ces liens soient associés à une même ancre, de façon à ce que les utilisateurs de Google, à chaque fois qu'ils formulent une requête contenant le(s) même(s) mot(s) que ceux de l'ancre en question, tombent sur une liste de résultats dans laquelle figurera en première position la page vers laquelle les « bombes Google » ont pointé.

La plus célèbre « bombe » a eu lieu lorsque des opposants au président des Etats-Unis George W. Bush eurent massivement créé des liens sur l'ancre « *miserable failure* » pointant vers la page officielle de George W. Bush sur le site de la Maison Blanche. Lorsqu'une requête était effectuée sur *Google* comprenant les termes « *miserable failure* », la page officielle de George W. Bush remontait ainsi en tête des listes de résultats générées par le moteur. Cela a constitué pour des internautes militants une manière de s'approprier l'algorithme du moteur afin de protester en ligne [Tatum, 2005].

Une étude a montré que de telles « bombes » pouvaient rester actives plusieurs mois et même plusieurs années [Bar-Ilan, 2007]. Au début, Google ne voulait pas intervenir, et les opinions différaient quant à la nécessité de son intervention [Grimmelmann, 2007b, p. 46]. Puis, finalement, des modifications de l'algorithme furent effectuées visant à diminuer l'impact de ce genre de stratégie [Cutts, 2007].

Certains référenceurs estiment que s'ils ne faisaient pas de SEO en allant parfois dans le sens contraire ce qui est préconisé par Google, les sites dont le contenu est de piètre qualité mais dont les éditeurs pratiquent le *black hat SEO* réussiraient à tromper l'algorithme et à se retrouver en haut des listes de résultats. Désobéir au code de conduite publié par Google pourrait donc être dans certains cas bénéfique à l'internaute et à Google.

Cette observation est confirmée par les travaux des économistes Ron Berman et Zsolt Katona selon lesquels le SEO peut profiter à un moteur de recherche et à ses utilisateurs dès lors que la valeur qu'un éditeur attribue au trafic dirigé sur son site est proche de la valeur que les internautes attribuent au contenu qu'ils y trouvent [Berman et Katona, 2013]. Dans le cas contraire, les économistes affirment que les pratiques d'optimisation pourraient pénaliser aussi bien le propriétaire du moteur de recherche que ses utilisateurs. Ainsi, dans le cas où un éditeur produirait un contenu valorisé par les internautes et pratiquerait le SEO pour remonter dans les résultats des moteurs, y compris s'il s'adonnait pour cela à des méthodes estampillées « *black hat* », cela pourrait avoir un effet positif pour l'ensemble des acteurs engagés.

Ajustements / réajustements

Les concepteurs de moteurs de recherche re-paramètrent leurs algorithmes en fonction des agissements qu'ils jugent néfastes à la pertinence des listes de résultats. Une fois qu'ils ont changé les critères et/ou les pondérations d'un algorithme, de nombreux éditeurs essayent de deviner ce qu'ils ont fait en testant certaines variables, une à une, grâce à des méthodes dites de « *A/B testing* », pour voir comment le moteur réagit et tenter de comprendre comme il est possible de remonter dans les classements.

Puisque l'algorithme du moteur dominant *Google* change régulièrement, la pratique du SEO exige que soient effectués une veille constante et des tests réguliers. Un jeu d'ajustement

et de réajustements se met en place entre les concepteurs des moteurs de recherche et les éditeurs de contenus, la tactique des premiers changeant en fonction de la stratégie des seconds, et vice et versa.

Depuis le début des années 2000, Google a procédé à de nombreux changements de son algorithme de classement. Jusqu'en 2009, ceux-là avaient lieu une fois par mois, le jour de la « *Google Dance* ». Tous les éditeurs s'empressaient de constater les effets des changements algorithmiques sur le classement de leurs documents. Depuis 2009, les changements ont lieu constamment et de nombreux ajustements ne font l'objet d'aucune communication de la part de Google. Seuls les changement majeurs sont accompagnés par des explications et des recommandations. Ces changements aux noms drolatiques — dont les plus célèbres sont *Esmeralda* (juin 2003), *Bourbon* (mai 2005), *Big Daddy* (février 2006), *Caffeine* (août 2009), *Panda* (février 2011) et *Pingouin* (avril 2012) — visent à améliorer la pertinence des résultats et à reléguer au fond des listes les « *black hats* ».

A la suite de la modification « *Panda* », par exemple, les fermes de contenus — c'est-à-dire les sites dont les contenus ont une très faible valeur ajoutée mais dont les pages sont nombreuses, conçues expressément pour occuper l'espace dans les résultats des moteurs de recherche en répondant à un grand nombre de requêtes et en maximisant l'affichage de bannières publicitaires — ont largement perdu en visibilité sur Google. A l'inverse, des sites plus qualitatifs comme ceux du *New York Times* et de *Yahoo news* ont bénéficié de la modification, tandis que certains sites qui n'avaient pourtant rien de *black hat*, comme celui du *British Medical Journal*, ont vu leur trafic baisser.

Finalement, l'algorithme est le résultat d'une influence mutuelle entre les actions des concepteurs du moteur de recherche et les actions des éditeurs de contenus. Le moteur de recherche est l'endroit d'une tension entre les intérêts de ces différents acteurs, leurs objectifs et leurs capacités d'action, dont dépendent *in fine* les modalités de fonctionnement du moteur et le classement des documents les uns par rapport aux autres.