Influence des algorithmes des outils de recherche sur les requetes d'une veille : exemple de la navigation anonyme.

Philippe Pinczon Du Sel

INFLUENCE DES ALGORITHMES DES OUTILS DE RECHERCHE SUR LES REQUETES D'UNE VEILLE : EXEMPLE DE LA NAVIGATION ANONYME

Philippe PINCZON du SEL (*)
pinczon@univ-tln.fr

(*) Laboratoire I3M, Université du Sud Toulon-Var, 83957 La Garde Cedex, France.

Mots clefs :
Veille stratégique sur Internet, aide à la prise de décision, algorithme user centric, géolocalisation, classement de requêtes, pertinence de l'information, moteurs de recherche, navigation anonyme
Keywords:
Competitive watch, decision-making, user centric algorithm, geolocalisation, queries rating, information relevance, search engines, anonymous web surfing
Palabras clave :
Vigilancia estratégica, ayuda en la toma de decisiones, algoritmo user centric, geolocalización, clasificación de solicitudes, pertinencia de la información, motores de búsqueda, navegación anónima

Résumé

Les moteurs de recherche s'appuient, en autres, sur des algorithmes dits « User Centric » ainsi que de géolocalisation afin d'affiner la pertinence de leurs résultats : les habitudes de navigation des internautes et leur lieu de connexion sont pris en compte lors du classement de l'information.
Les campagnes de veille sur Internet sont principalement menées avec l'aide des moteurs de recherche, parmi lesquels certains utilisent ces algorithmes. Les personnes en charge de ces recherches peuvent donc à priori influencer de leurs habitudes de navigation et de leur situation géographique les résultats des requêtes : elles auraient été différentes selon la personne et le lieu.
Ces résultats aidant, par la suite, à la prise de décisions parfois stratégiques au sein des entreprises, nous pouvons penser que les veilleurs influent indirectement ces décisions.
L'utilisation des méthodes de navigation anonyme pour des recherches d'informations sur Internet permettrait d'inhiber l'action de ces algorithmes tout en pérennisant l'utilisation de ces moteurs de recherche : en effet, les outils de navigation anonyme bloquent, entre autres, les cookies et les adresse IP des ordinateurs, empêchant ainsi aux moteurs de recherche de personnaliser le classement des résultats des requêtes.

1 Introduction

Les moteurs de recherche s'appuient sur deux métiers pour fournir aux internautes des réponses à leurs recherches : ils indexent (collectent et stockent) d'une part l'ensemble des pages présentes sur le web, puis d'autre part les classent selon des critères de pertinence afin de présenter des résultats cohérents aux requêtes qui leur sont soumises.

Les critères de pertinence évoluent avec le temps et varient selon les outils. Ainsi la pertinence d'une page peut dépendre de son contenu : on parle alors d'algorithmes « content centric » selon lesquels la fréquence d'apparition d'un mot dans la page, sa rareté, sa typographie ou sa présence dans les balises méta en font un critère essentiel. La pertinence d'une page peut également dépendre de sa popularité : les algorithmes « link centric » estiment qu'une page est populaire si elle est citée par d'autres pages, qu'elles soient ou non issues de sites de catégories similaires (algorithmes relationnels et contextuels). Enfin, d'autres algorithmes intègrent les habitudes des internautes en lisant les fichiers cookies des postes informatiques (algorithmes « user centric ») ainsi que leur situation géographique en analysant les fichiers de connexion ou les paramètres des navigateurs utilisés.

Ces deux derniers algorithmes sont donc dépendants d'informations que les internautes leur fourniraient. La pertinence, et par extension le classement d'une page, ne dépendrait plus d'elle-même ou de ses relations avec d'autres pages mais de son adéquation avec l'internaute. La pertinence d'une page ne serait donc plus figée mais aléatoire, au hasard de la personnalité de l'internaute et de sa situation géographique au moment du questionnement. La pertinence d'une page ne serait plus un critère interne et contrôlable, mais dépendrait désormais d'éléments externes non maîtrisables.

En partant de cette hypothèse et au travers d'une expérience, cet article propose de vérifier l'influence effective de ces algorithmes sur les résultats de requêtes posées à un moteur de recherche et par voie de conséquence l'influence que peuvent avoir des personnes chargées d'une veille sur Internet sur ces résultats.

2 Origines de la recherche

2.1 Les risques liés à l'utilisation de l'Internet

Jakobiak [1] notait en introduction de son chapitre dédié à la sécurité de l'information en Intelligence Economique que le spécialiste de la veille stratégique ou de l'intelligence économique prend conscience, en surveillant les autres, les concurrents, de la nécessité d'être discret, d'être prudent, de se protéger contre les risques et menaces diverses qui pourraient porter atteinte à son patrimoine informatif. Il met l'accent sur les différents types de menaces, dresse des portraits d'attaquants et liste une série d'attaques possibles avant de préciser que la méfiance des spécialistes de l'intelligence économique vis-à-vis de l'Internet tient à un certain nombre de risques du système : confidentialité difficile à respecter, désinformation possible, modification de documents, perte totale ou partielle de documents en cours de transfert, virus…

Par ailleurs, Revelli [2] identifie quatre grands thèmes de sécurité sur Internet : les attaques, les virus, le cryptage et l'espionnage. Si les deux premières peuvent être associées au piratage en général, les dernières correspondent bien à la nouvelle forme de risque lié à Internet. Ainsi, préconise-t-il de faire attention aux interventions dans les forums de discussion, d'éviter d'utiliser le réseau d'une entreprise pour des recherches confidentielles et de désactiver les cookies à tout moment.

Romagni & al. [3] nous mettent en garde à propos des problèmes de sécurité liés à l'utilisation de l'Internet dans un système de veille : dans quelques années, les technologies de tracing permettront aux serveurs Internet de garder votre identifiant et ainsi retrouver votre identité…

Enfin, le Conseil d'Etat [4] souligne que les réseaux apportent un enrichissement formidable des possibilités de faire et de mal faire, grâce notamment à l'interactivité, et surtout à la possibilité d'une circulation internationale des informations. Deux nouveaux types d'atteintes apparaissent selon les différents types de services proposés à l'utilisateur : soit les renseignements sont collectés sur les individus à travers des traitements visibles, soit une collecte d'informations est réalisée à l'insu de ceux-ci, au moyen de techniques mises en œuvre de façon cachée.

2.2 Identification des risques

Une première distinction est ici faite entre le risque de piratage proprement dit et la nécessité de se protéger contre les nouvelles technologies de l'Internet (NTI), puis une seconde distinction est proposée au sein même des NTI entre les informations diffusées en connaissance de cause par les internautes et celles qui leur sont soutirées sans qu'ils en soient avertis.

Divers éléments sur les internautes peuvent être recueillis de façon transparente sur le réseau : pratiques observées dans les forums de discussion, exploitation des messages électroniques et des annuaires. Mais au-delà des données nominatives circulant sur le réseau, il en est d'autres que l'utilisateur ne peut directement appréhender, alors que leur valeur informationnelle et les risques qu'elles représentent pour la vie privée des personnes sont importants : fichiers logs et cookies (Conseil d'Etat [4]).

Les premiers servent à établir la communication entre ordinateurs distants. Les informations qui y sont stockées concernent d'une part les adresses des machines du réseau, dites adresses IP et en particulier celles de l'émetteur d'un message et de son destinataire, adresses auxquelles sont associées la date et l'heure de la connexion, des informations techniques caractérisant le type d'usage (accès au web, messagerie…) et d'autre part la requête (page du site que l'utilisateur veut visiter) ou le message proprement dit. Ces données sont collectées automatiquement par les fournisseurs d'accès et consignées dans un fichier dénommé « fichier log » (Conseil d'Etat [4]).

Le cookie est un petit fichier texte émis par un serveur consulté par un utilisateur et enregistré sur le disque dur de celui-ci. Il comporte, en général, une date de validité et peut contenir l'information qu'aura souhaité y inclure le site visité (Conseil d'Etat [4]) : ce peuvent être les dates et heures des visites (ou un compteur de visites), un numéro client, des réponses à un questionnaire, des mots-clés tapés dans le champ de recherche du site, des préférences ou le choix d'affichage et de navigation... Dans le cas d'un site web d'achats en ligne, les produits placés dans le « caddie » le sont en fait dans le cookie du site. En somme, le principe du cookie est purement marketing et a pour but de savoir si c'est la première fois que vous vous connectez à un certain site ou si au contraire vous y êtes déjà passé (Revelli [2]) : il permet au responsable du site de mémoriser les précédentes consultations du site par l'internaute afin, soit de faciliter l'ergonomie de la visite, soit d'adapter les pages du site au « profil » de l'internaute tel qu'il est précisément déduit des « traces » conservées lors des précédentes visites (Conseil d'Etat [4]).

Les internautes n'ont aucun contrôle sur ces fichiers logs, et les supprimer entraînerait automatiquement la perte de communication avec l'Internet. Par ailleurs, refuser d'accepter des cookies entrave considérablement la navigation sur les sites web : certains d'entre eux refusent en effet leur accès dès lors que les cookies ont été désactivés.

2.3 Une solution : la navigation anonyme

Une solution existerait dans la navigation anonyme : cette technique consiste à se connecter à des proxies, généralement situés à l'étranger, qui sont configurés de telle sorte à ne pas transmettre les informations telles que les adresses IP ou les cookies d'un internaute vers le site visité, mais fourni en place leurs propres identités (figure 1). L'internaute peut alors naviguer « anonymement » puisque ses paramètres ne sont pas divulgués au-delà de ces proxies, également appelés « anonymizers ». Ils servent d'intermédiaires entre les internautes et Internet dans sa globalité.

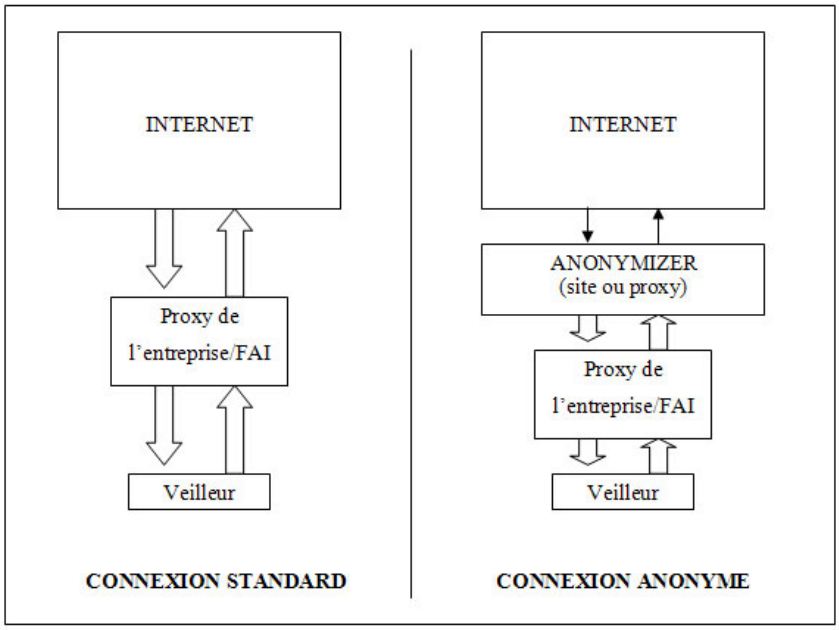

Figure 1 : Principe d'une connexion anonyme sur Internet passant par un anonymizer

3 Application de l'anonymat en ligne dans un système de veille

3.1 L'hypothèse

Les algorithmes « user centric » et de géolocalisation se basent sur deux informations présentes dans les disques durs des ordinateurs : les cookies pour l'un et l'adresse IP, donc les fichiers logs, pour l'autre.

Ces deux informations, nous l'avons vu, sont nécessaires pour une bonne navigation sur Internet, il est donc difficile de les inhiber directement depuis les ordinateurs. En revanche, les proxies « anonymizers » le permettent.

L'hypothèse que nous proposons de vérifier et dont nous distinguons deux aspects, est la suivante :

- D'une part, si ces outils garantissent réellement l'anonymat sur Internet, ils devraient permettre de contourner l'influence des algorithmes des moteurs de recherche « user centric » et de géolocalisation. Les résultats obtenus pour une requête donnée devraient donc être différents de ceux que l'on obtient de façon standard.

- D'autre part, ces outils devraient nous permettre d'accéder à des pages de sites Web qui seraient inaccessibles pour diverses raisons : le site d'enchères en ligne eBay bloque par exemple pour des raisons juridiques l'accès à certains objets aux enchérisseurs Français.

3.2 L'expérimentation

Une première expérience visant à vérifier cela a été menée au sein de l'Université du Sud Toulon Var. Afin de réduire autant que possible le risque d'erreur pouvant entraîner de faux résultats, nous avons décidé de suivre un protocole strict qui prévoit d'utiliser de façon simultanée :

- Deux postes informatiques identiques, de type PC et possédant les mêmes caractéristiques matérielles (hardware) ;
- Des navigateurs et des systèmes d'exploitations identiques (software) ;

- Une connexion Internet au travers du proxy de l'Université du Sud Toulon Var ;
- Des disques durs formatés pour l'occasion.

Ainsi, les caractéristiques d'environnement des machines sont identiques et l'adresse IP diffusée sur Internet est la même pour les deux machines ; les disques durs ayant été préalablement formatés, aucun fichier ou programme commercial (cookie, spyware) ne peut influencer les résultats des requêtes.

De plus, l'utilisation simultanée des deux machines réduit de manière significative les marges d'erreurs liées à la temporalité des résultats des requêtes.

Par ailleurs, en ce qui concerne le choix des sites web visités et des requêtes tapées :

- Nous avons opté pour le moteur de recherche Google.com, version anglophone afin d'obtenir un maximum de résultats ;
- Nous avons opté pour le site eBay.com, version anglophone également, pour les raisons que nous avons décrites auparavant : le site d'enchères en ligne interdit l'accès à certaines de ses pages aux enchérisseurs Français ;
- Nous avons opté pour l'outil de navigation en ligne guardster.com, choisi parmi d'autres pour sa disponibilité et son efficacité ;
- Les mots-clés tapés dans le moteur de recherche sont des noms d'entreprises internationales et nationales soumises à une concurrence internationale ainsi qu'un terme générique commercial. Les premiers nous permettent de simuler une veille technologique, le second nous permettra de vérifier l'influence réelle des algorithmes de pertinence ;
- Les deux mots-clés utilisés sur le site d'enchères en ligne n'appartiennent à aucune catégorie interdite, mais le premier a toutefois été choisi pour ses possibles relations avec elles.

En choisissant d'utiliser uniquement la langue anglaise lors de cette expérimentation, nous élargissons non seulement notre champ de recherche, mais cela nous permet d'obtenir des résultats provenant de pages web et de requêtes strictement comparables : une même requête posée sur Google France par exemple ne donne pas les mêmes résultats que sur le site principal anglophone.

Afin que l'expérience soit la plus juste possible, nous devons faire en sorte d'obtenir, sur le moteur de recherche Google, la même interface de requête en modes de navigation standard et anonyme : il est donc important préciser dans le premier cas que les recherches se font en anglais (*« google in english »*). En revanche, en mode anonyme, l'interface anglophone est immédiatement proposée : nous ne sommes plus identifiés comme étant Francophones.

La série de mots-clés est ensuite tapée simultanément sur les deux postes, puis nous notons, pour chaque requête, le nombre de résultats ainsi que les noms et URL des soixante premiers sites web, soit ceux des trois premières pages.

L'interface du site ebay.com est très différente que l'on soit en mode de navigation standard ou anonyme, bien que dans les deux cas nous obtenons effectivement des interfaces strictement anglophones.

Nous procédons de la même façon que précédemment. Les mots-clés sont tapés simultanément, nous notons le nombre d'objets proposés puis nous vérifions leur accessibilité : en cliquant sur chacun des cinquante liens des premières pages de résultats nous vérifions si nous avons bien accès aux objets listés.

	NAVIGATION STANDARD	NAVIGATION ANONYME
MOTEUR DE RECHERCHE : GOOGLE	**http://www.google.com** **option: « google in english »**	**http://proxy.guardster.com/cgi-bin/nph-proxy.cgi** **requête: http://www.google.com**
Requêtes	**Nombre de réponses**	**Nombre de réponses**
« boeing »	**4 490 000**	**7 390 000 (+ 64,5%)**
« dassault aviation »	**153 000**	**183 000 (+ 19,6%)**
« cea »	**2 810 000**	**2 930 000 (+ 4,2%)**
« cellular ringtones »	**3 140 000**	**3 270 000 (+ 4,1%)**
SITE COMMERCIAL : EBAY	**http://www.ebay.com**	**http://proxy.guardster.com/cgi-bin/nph-proxy.cgi** **requête: http://www.ebay.com**
Requête	**Nombre d'objets**	**Nombre d'objets**
« army medals »	**468** **accessibilité : 33/50 (66%)**	**469 (+ 0,02%)** **accessibilité : 50/50 (100%)**
« cosmetics »	**6 552** **accessibilité : 50/50 (100%)**	**6 552 (idem)** **accessibilité : 50/50 (100%)**

Tableau 1 : Tableau comparatif du nombre de réponses aux requêtes tapées sur les sites Google.com et eBay.com en modes de navigation standard et anonyme

Les résultats de l'expérimentation sont présentés dans le tableau 1. Sa lecture nous apprend plusieurs choses :

- Nous obtenons plus de réponses à nos requêtes tapées dans le moteur de recherche lorsque nous naviguons anonymement que lors d'une recherche habituelle ;
- Cette différence peut atteindre le double de résultats pour certains mots-clés (+ 64,5% pour la requête « boeing ») ;
- Nous avons effectivement accès aux objets traditionnellement bloqués par le site d'enchères en ligne.

Néanmoins, cela nous amène à nous poser des questions :

Les informations sont-elles réellement cachées ou est-ce un problème purement technique ?

Dans le cas du site d'enchères en ligne nous savons que certaines informations sont cachées pour des raisons juridiques ; en revanche en ce qui concerne le moteur de recherche, la différence de résultats peut être le fait de l'inhibition des algorithmes, mais cela peut également être dû au mode de fonctionnement du moteur de recherche Google, comme par exemple le fait d'utiliser des serveurs-mirroirs : les requêtes tapées lors de la connexion anonyme auraient alors été dirigées vers un autre serveur, estimé comme étant plus proche de notre nouvelle identité, et plus récemment mis-à-jour que celui qui a fourni les réponses en mode standard.

Au regard des résultats sur les sites Français « dassault aviation » et « cea » ainsi que sur le mot-clé à caractère commercial « cellular ringtones », nous penchons vers une de ces deux dernières hypothèses.

En effet, l'écart est relativement peu important et peut aisément être justifié par l'absence des algorithmes « user centric » et de géolocalisation ou de la redirection vers un autre serveur-mirroir.

En revanche, la requête « boeing » est surprenante par son écart de résultat. Il semblerait que d'autres paramètres aient été pris en compte - ou non - lors de l'affichage des résultats en mode anonyme : les hypothèses citées précédemment concernant la non prise en compte des algorithmes ou d'une éventuelle redirection vers d'autres serveurs ne permettent plus de justifier une telle différence.

Cette expérience a donc démontré que les résultats des requêtes sont quantitativement supérieures lors d'une recherche anonyme, mais le sont-elles également qualitativement parlant ?

Plus de résultats n'impliquent pas automatiquement de meilleurs résultats : par exemple, les réponses supplémentaires obtenues en mode de navigation anonyme ont peut-être été omises lors de la recherche en mode standard grâce, justement, à l'action de l'algorithme « user centric ». Néanmoins, un rapide coup d'œil sur les pages de résultats permet de constater que dès les dix premières réponses, l'ordre de liens est modifié, certaines réponses apparaissent et d'autres disparaissent. Dès la troisième page, soit vers la cinquantième réponse, les résultats n'ont plus aucune concordance.

Il sera intéressant lors de la suite de l'expérimentation d'analyser plus finement les listes des réponses afin d'essayer de déterminer à la fois pour quelle raison technique elles sont différentes et laquelle propose une meilleure pertinence des résultats. Il faudra toutefois tenir compte du fait que les internautes n'ont accès qu'aux mille premières réponses d'une requête, et que d'une manière générale ils lisent rarement les résultats au-delà des deux premières pages.

3.3 Application de l'expérimentation à la veille

Les moteurs de recherche évoluent et tendent vers une amélioration de la pertinence des résultats proposés aux internautes. Or, cette pertinence est toute relative : elle est basée sur l'hypothèse que les résultats sont d'autant plus pertinents pour un internaute s'ils prennent en compte ses goûts, ses habitudes de navigation et sa situation géographique.

Nous pensons que la recherche d'informations sur Internet dans le cadre d'une veille technologique doit s'affranchir de ces évolutions : pour atteindre un niveau optimal d'efficacité, une veille sectorielle sur Internet doit demeurer impersonnelle, les résultats des recherches sont faussés si on y intègre la localisation géographique et les goûts personnels des veilleurs en charge de l'action.

Par ailleurs, la technique le permettant - le site d'enchères en ligne en est l'exemple -, nous pensons que des sociétés soumises à forte concurrence pourraient cacher ou modifier leurs pages selon l'origine du visiteur. Toutefois, l'expérimentation avec le site d'enchères en ligne n'est destinée qu'à démontrer l'efficacité de la navigation anonyme : dans ce cas précis, l'interdiction faite aux enchérisseurs Français est basée uniquement sur le paramétrage linguistique des navigateurs. Il est donc possible de contourner l'interdiction autrement qu'en passant par les services d'un site d'anonymat en ligne.

De la même manière, des sites web mal configurés pourraient involontairement empêcher leur accès à une catégorie d'internautes, empêchant ainsi toute récolte d'informations. Nous parlons ici d'information blanche, accessible à tous mais qui pourrait être cachée ou modifiée dans un but de désinformation par exemple.

La navigation anonyme peut être considérée comme une piste vers une solution qui permettrait d'éviter ces situations. Elle ouvrirait la voie vers l'obtention d'informations supplémentaires en ligne, en proposant à la fois des résultats différents aux requêtes posées aux moteurs de recherche mais aussi en permettant d'accéder aux pages qui seraient cachées pour des raisons purement techniques (nous ne parlons pas de piratage).

4 Conclusion

L’objectif de cette recherche vise à améliorer sensiblement la recherche d’informations sur Internet, notamment dans le cadre d’une veille sectorielle, en proposant d’une part d’obtenir des résultats de moteurs de recherche vierges de toute influence et d’autre part d’accéder à des pages, donc à de l’information, qui ne l’auraient pas été autrement.

Cette première expérience a permis de constater que la navigation anonyme influe effectivement sur les résultats des moteurs de recherche et sur la présentation des pages web lors de recherches d’information sur Internet. La suite de l’expérimentation permettra d’approfondir la recherche en répondant à d’autres questions, notamment si la pertinence des réponses et leur classement sont meilleurs en mode anonyme, si les algorithmes « user centric » et de géolocalisation, entre autres, expliquent à eux seuls la différence de résultats obtenue avec la requête sur le constructeur Boeing, ou s’il existe d’autres raisons pour expliquer ces différences de résultats. Par ailleurs, au-délà de ces problèmes techniques, la navigation anonyme pourrait-elle devenir un outil pour contrer la désinformation en ligne en permettant à une catégorie d’internautes d’accéder aux pages qui leur seraient interdites ou plutôt d’éviter celles qui leur seraient destinées ? Enfin, il sera intéressant d’analyser et de comprendre l’influence de la navigation anonyme sur le référencement des sites.

5 Bibliographie

[1] JACOBIAK F., L'intelligence économique en pratique, 2ème édition, Les éditions d'organisation, 2001

[2] REVELLI, C., Intelligence stratégique sur internet, Dunod, 2000

[3] ROMAGNI, P., et WILD, V., L’intelligence économique au service de l’entreprise, 1998

[4] CONSEIL D'ETAT, Internet et les réseaux numériques, La documentation Française, 1998