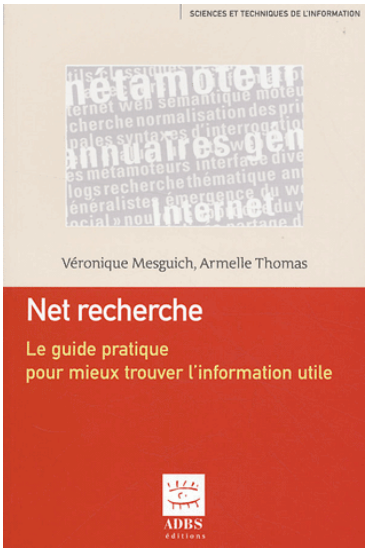
Net recherche : Comment comparer les moteurs ? Où trouver des statistiques sur les moteurs ? Peut-on garder un historique des recherches ? Retrouver des pages disparues ?

Retour au sommaire de la lettre

Domaine :	**Recherche**	Référencement
Niveau :	**Pour tous**	Avancé

Nous vous proposons dans cet article deux "fiches pratiques" extraites du livre "Net recherche" qui vient de sortir aux éditions ADBS. Elles nous décrivent les différentes façons de comparer les moteurs de recherche ainsi que les sources identifiées pour trouver des informations à leur sujet. La seconde fiche nous parle des différentes façons de rechercher une information ancienne, ayant disparu du Web au moment où l'on effectue une investigation numérique...

Le livre "***Net recherche 2009 : le guide pratique pour mieux trouver l'information utile et surveiller le web***" vient de paraître aux éditions ADBS (http://www.adbs.fr/net-recherche-2009-le-guide-pratique-pour-mieux-trouver-l-information-utile-et-surveiller-le-web-61812.htm).
En voici une courte présentation proposée par l'éditeur :



Sous l'apparente facilité d'utilisation des moteurs de recherche se cache en effet une réalité complexe, et le secret de la réussite d'une recherche ou d'une veille passe autant par la maîtrise des aspects techniques que par la capacité à évaluer et sélectionner les sources pertinentes.
En cette troisième édition profondément renouvelée, développée et mise à jour, Net recherche vise à offrir à toute personne amenée à effectuer des recherches sur Internet un panorama des outils et méthodes existant à ce jour, en intégrant les dispositifs qui permettent de surveiller le web à moindre coût.
Cet ouvrage s'adresse notamment aux professionnels de l'information (documentalistes, bibliothécaires, veilleurs), aux enseignants, chercheurs et étudiants, et à tous les autres "travailleurs du savoir" confrontés à la complexité croissante et à l'inflation de l'information en ligne. Ils y découvriront des conseils méthodologiques mais aussi nombre de "trucs et astuces" destinés à optimiser le processus de recherche ou de veille, des informations précises sur les évolutions en cours, ainsi que des présentations d'outils et des listes d'adresses utiles.

Les auteurs ont demandé à Olivier Andrieu, éditeur du site Abondance.com d'en signer la préface. En accord avec les auteurs et l'éditeur, nous vous présenterons, pendant quelques mois, quelques "fiches pratiques" présentes dans ce livre. Nous espérons qu'elles vous permettront de mieux chercher l'information sur le Web... Après "*Comment choisir ses mots-clés ?*", "*Comment effectuer un sourcing de qualité*" il y a deux mois puis "*Comment trouver des articles de presse*" le mois dernier, voici deux nouveaux articles extraits de cet ouvrage...

Fiche 4. Comment comparer les moteurs ? Où trouver des statistiques sur les moteurs ?

Le moteur idéal n'existe malheureusement pas... Chaque moteur possède ses points forts et ses points faibles, et il est difficile d'établir des classements totalement objectifs. Nous pouvons toutefois établir des critères généraux de comparaison, étudier plus en détail les fameux algorithmes de pertinence, et aussi lister les outils qui permettent de comparer, visuellement ou non, les résultats de différents moteurs pour une requête donnée.

En comparant les moteurs choisis selon différents critères

Différents critères peuvent servir à évaluer la qualité d'un moteur de recherche, et à comparer les moteurs entre eux.
- Provenance du (ou des) index, taille (approximative !) de l'index, ressources prises en compte (actualités, fichiers images, audios, vidéos, forums, etc.).
- Délai moyen de rafraîchissement et conditions de mise à jour.
- Mode d'indexation et traitement éventuel des ressources : linguistique, statistique, *parsing* (extraction des éléments signifiants).
- Options de recherche simple et avancée, aide à la reformulation des questions, tolérance aux fautes d'orthographe.
- Critères déterminants pour le classement des résultats (tri de pertinence).
- Pertinence des résultats (avec toute la difficulté d'appréciation liée à ce critère !)
- Présentation des résultats : informations disponibles, source du résumé, datation des résultats, regroupement des pages d'un même site (cluster), vignettes des pages résultats, mise en exergue des mots-clés sur la page et des pages avec photos ou fichiers spécifiques, archive de la page, cartographie, autres fonctionnalités d'aide à l'usager (dictionnaire, traduction, actualités), etc.
- Possibilité d'une personnalisation, de conservation de l'historique.
- Rythme d'innovations.
- Critères subjectifs : interface de consultation, adéquation aux types de recherche effectués.
- Critères de marché : part de marché, cibles principales, types d'usagers, etc.

En utilisant un métamoteur dédié

Les moteurs majeurs du web délivrent-ils tous à peu près les mêmes réponses à une recherche donnée ? Cette question est importante, puisque le nombre de résultats devient souvent de plus en plus gigantesque, et impose souvent des stratégies complémentaires ou alternatives.

Contrairement aux idées reçues, les premiers résultats s'avèrent fort différents d'un moteur à l'autre. Une étude conduite en juillet 2005 par le métamoteur Dogpile, en collaboration avec des chercheurs des universités de Pittsburgh et de Pennsylvanie, montrait que le taux de recouvrement était alors très faible sur les dix premiers résultats : seulement 1,1 % des liens proposés apparaissaient communs aux quatre moteurs testés (Google, Yahoo!, Live Search, Ask), 89,4 % étant uniques à un seul moteur et 11,4 % étant proposés par deux moteurs.

En avril 2007, une nouvelle étude réalisée sur les mêmes moteurs avec des chercheurs de l'Université de Pennsylvanie et de celle de Queensland (plus de 19 000 requêtes testées cette fois, pour 12 000 dans le cadre de l'étude précédente) donne un taux de recouvrement encore plus bas sur les dix premiers résultats, de l'ordre de 0,6 % ! En d'autres termes, la majorité des résultats de la première page sont spécifiques à chaque moteur. Notons que le pourcentage de résultats totaux partagés par deux moteurs serait de 8,9 % (http://www.infospaceinc.com/onlineprod/Overlap-DifferentEnginesDifferentResults.pdf).

Pour comparer plus facilement les différents résultats sur une requête, on pourra exploiter les métamoteurs qui ont fait de cette fonction leur cœur de métier. Ainsi, **Grab All** (http://www.graball.com/) permet de se rendre compte d'un coup d'œil des choix de premiers résultats par moteur.

Twingine (http://www.twingine.com/) fait de même, mais uniquement avec Google et Yahoo!.

MywebSearch (http://search.mywebsearch.com/mywebsearch/default.jhtml), créé par Ask, est très pratique pour comparer les résultats de Google, Yahoo! et Ask, grâce à son système d'onglets.

Thumbshots (http://www.thumbshots.org/Products/Thumbshots/Ranking.aspx) développe une interface de visualisation très intéressante des recouvrements entre deux moteurs parmi une liste intégrant notamment Google, Yahoo! et Live Search. Il utilise sa technologie d'"aperçus de pages web" pour faciliter l'exploration des liens.

Releton (http://www.releton.com/) permet non seulement d'effectuer une recherche simultanément sur Yahoo! et Google mais aussi d'augmenter, à l'aide d'un curseur, la part des résultats venant de l'un ou l'autre de ces moteurs ; par défaut, le curseur est sur une position 50/50, offrant une parité entre les résultats des deux moteurs.

Jux2 (http://jux2.com/), autre métamoteur intéressant dans cette optique, ne proposait plus ses modes avancés de comparaison en février 2009 (par exemple, réponse à la question : *quels sont les résultats de Google qui ne sont pas dans Yahoo! ?*).

En étudiant les statistiques sur les moteurs

Plusieurs sociétés sont spécialisées dans la mesure d'audience des sites web et l'analyse de trafic et étudient notamment les "parts de marché" des différents outils de recherche du web. Voici les principaux sites à consulter.

En France :
- Baromètre Xiti : http://barometre.secrets2moteurs.com/
- Baromètre Adoc : http://www.barometre.adoc.fr/

Des sociétés étudient plus généralement les sites commerciaux, comme Mediametrie-eStat ou Wysistat, mais peuvent parfois donner des informations utiles sur les moteurs.

À l'international :
- Comscore : http://www.comscore.com/
- Nielsen Online : http://www.nielsen-online.com/
- Hitwise : http://www.hitwise.com/
- OneStat : http://www.onestat.com/

Fiche 5. Peut-on garder un historique des recherches ? Retrouver des pages disparues ?

Il peut être intéressant de mémoriser une équation de recherche complexe, afin de la réactualiser régulièrement dans une optique de veille ou tout simplement pour ne pas avoir à ressaisir les termes. Nous avons vu au chapitre 5 plusieurs formules de surveillance des mots-clés, et notamment le service Google Alert. Nous allons, en guise de complément, étudier ici le moyen de sauvegarder des stratégies de recherche.

Un outil comme **Copernic** (http://www.copernic.com/fr/), par exemple, permet de conserver les stratégies de recherche et les résultats, en les classant éventuellement dans des dossiers thématiques, avec une fonction de veille automatisée disponible dans la version payante de l'outil.

Les recherches effectuées *via* les moteurs comme Google ou Yahoo!, du moins les plus récentes, peuvent toujours être retrouvées dans l'historique du navigateur... À noter que Google a lancé au printemps 2005 sur son site américain une option baptisée ***"Google Search History"*** (http://www.google.com/psearch) qui permet de mémoriser l'historique des recherches. Cette option, aujourd'hui disponible en français, nécessite la création d'un compte de messagerie Gmail, mais on peut y accéder également à partir de toute adresse électronique. Elle s'inscrit ainsi dans les éléments de personnalisation, et donc de fidélisation des internautes. On peut ainsi, grâce à un calendrier, retrouver les requêtes effectuées tel ou tel jour. On peut également générer des statistiques concernant le nombre de recherches journalières, sous forme d'histogramme.

L'option similaire de Yahoo!, "My web search", devait disparaître en mars 2009 et devrait être remplacée par l'outil **Search Pad** (http://help.yahoo.com/l/us/yahoo/search/searchpad/), destiné à faciliter la mémorisation des recherches et des liens visités et l'ajout de commentaires.

Enfin, les éléments de recherche peuvent être mémorisés dans les formulaires des différents moteurs. Pour activer cette fonction, il suffit de paramétrer le navigateur. Sur Internet Explorer, dans le menu Outils, cliquer sur *Options Internet*, puis *Contenu*, *Informations personnelles/Saisie semi-automatique*, et cocher *Formulaires*. On peut ensuite effacer les éléments indésirables directement dans le formulaire. Dans Firefox, choisir *Options, Vie privée*, et cocher "*Se souvenir des informations saisies dans les formulaires et la barre de recherche*".

Où trouver des pages disparues ou modifiées ?

Si la page a été modifiée très récemment, elle sera peut-être encore disponible dans le "cache" proposé par de nombreux moteurs (dont Google, Yahoo! et Live) : il s'agit d'une copie de la page telle qu'elle se présentait lors du dernier passage du robot. Aucune page web ne peut prétendre à l'éternité. Pour autant, le service **Way Back Machine** (http://www.archive.org/web/web.php), proposé par l'association The Internet Archive (association à but non lucratif sous l'égide du californien Brewster Kahle qui reçoit des donations et soutiens de différents acteurs), est très impressionnant : on peut visualiser un site tel qu'il était à différentes dates depuis 1996, et même suivre des liens sur ces archives. Il suffit de taper l'adresse url d'une page (cela peut être une page d'accueil, ou bien une page donnée à l'intérieur d'un site). On obtient une liste de dates, correspondant à des versions antérieures de la page "capturée" à des moments divers, de façon aléatoire. Remarque : un embargo de six mois, voire un an, est pratiqué sur les versions les plus récentes.

À noter : la **Bibliothèque nationale de France** compte parmi ses missions le dépôt légal des sites publics français. Des expériences ont déjà été menées sur certains sites officiels (http://blog.bnf.fr/lecteurs/index.php/2009/06/23/les-archives-de-l-internet-francais-a-la-bnf-de-1996-a-aujourd-hui/). En 2007, la BnF s'est livrée à une collecte des sites électoraux. Les sites liés à l'élection présidentielle et aux élections législatives de 2007 ont été capturés sur une période de dix mois, d'octobre 2006 à juillet 2007. Le projet a permis la sauvegarde de plus de 5 800 sites ou parties de sites, à des fréquences régulières. Tous les types d'acteurs du débat politique sur la Toile ont été représentés : sites de candidats, de partis ou d'organisations de soutien, mais aussi blogs de militants, observatoires de la "Net-politique" ou presse en ligne.
La collection constituée représente un ensemble de 63 millions de fichiers, soit 3,4 téraoctets de données !

De son côté, l'**Initiative pour les archives ouvertes (OAI :** http://edutice.archives-ouvertes.fr/**)**, lancée en 2000, s'intéresse à l'ensemble des activités liées à l'archivage des publications scientifiques. On trouvera sur le site openarchives.org des outils et méthodes permettant l'interopérabilité de ces archives, en s'appuyant sur les métadonnées au format Dublin Core.

Enfin, on peut parcourir une sorte de "cimetière" de sites disparus, ou n'étant plus maintenus, sur le site Ghost Sites (http://www.disobey.com/ghostsites/)...

Véronique Mesguich et Armelle Thomas
Auteurs du livre "Net recherche 2009 : le guide pratique pour mieux trouver l'information utile et surveiller le web". Extraits du chapitre 7.

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/07/net-recherche-comment-comparer-les.html