

Panorama des différents moteurs de recherche

Février 2009

Définition

"Un moteur de recherche est une machine spécifique (matérielle et logicielle) chargée d'indexer des pages web afin de permettre une recherche à l'aide de mots-clés dans un formulaire de recherche"
L'indexation du web se fait régulièrement grâce à des robots (*spiders* ou *crawlers*) qui parcourent les pages web de lien en lien et les stockent dans leur index. Cet index est interrogé par l'internaute avec des mots clefs libres dans une interface de recherche. Le moteur interroge alors son index et sélectionne les pages web qui contiennent les mots clefs (recherche en plein texte). L'affichage de résultats se fait selon un classement propre à chaque moteur en fonction de certains critères de pertinence que l'on appelle *page rank* (occurrence des mots dans la page et dans le titre, popularité, mise à jour régulière...).

La plupart des moteurs de recherche peuvent être ajoutés dans la barre d'outils de votre navigateur internet.

I. Les moteurs généralistes

Les moteurs de recherche généralistes classiques

Google : un des moteurs les plus complets par le nombre d'options innovantes² qu'il propose régulièrement et le plus performant par la qualité de ses recherches et la pertinence de ses résultats. Son index dépasse 40 milliards de pages et ses critères de classement des résultats sont les plus objectifs existants. Ce moteur est actuellement prédominant en France avec près de 90% des requêtes (63% dans le monde). Il est possible de chercher un document en fonction de son support (textes, images, vidéos), par typologie (« Actualités » pour chercher dans la presse en ligne, recherche de blogs, de livres avec « Google Books »). Google propose l'alternative d'une recherche par **annuaire** (qui utilise DMOz qui est un annuaire contributif et bénévole).

Yahoo : Moteur concurrent de Google (12% des requêtes dans le monde), il offre sensiblement la même qualité de réponse lors des requêtes simples et un index presque aussi conséquent, mais est moins pertinent dans le cas d'une requête complexe. L'index est sensiblement de la même taille que Google. Yahoo propose l'alternative d'une recherche par **annuaire**.

¹ Définition de Comment Ca Marche : <http://www.commentcamarche.net/contents/web/moteur-recherche.php3>

² *Googlepapes pour créer des sites, Blogger pour créer des blogs, gestion de son agenda en ligne, créer des groupes de discussion, Picasa pour stocker et retoucher des photos...*

Exalead est un moteur de recherche conçu en France, basé sur la spécificité du langage français et fonctionnant sur le clustering pour générer des termes associés, mais aussi pour catégoriser à partir d'une liste définie (index de 8 milliards de pages). Il permet de prévisualiser les pages web grâce à des vignettes et des fonctions avancées sont proposées pour affiner la recherche (termes associés, type de ressources, langue, annuaire...). Il permet de créer et gérer une sélection de sites favoris grâce à l'option « Ajouter aux raccourcis » (portail personnalisé du type Netvibes). Il est tout de même à noter que la place des liens sponsorisés réduit considérablement la visibilité de l'affichage des résultats. Possibilité de faire des recherches phonétiques, lorsque l'on ne connaît pas l'orthographe d'un mot.

Ask propose une interface moderne, avec possibilité de la personnaliser avec des skins. L'affichage des résultats sous forme de liste offre cependant plusieurs options : termes associés, visualisation de la page web, réponses flash (par exemple : définition du mot clef par Wikipédia, site officiel...), possibilité de cataloguer un site et le classer dans un dossier dans un espace personnalisé. L'algorithme Expert Rank du moteur Ask se base non seulement sur la popularité des liens, mais aussi sur l'identification des sujets (clusters) pour améliorer la pertinence des résultats.

Wikiasearch est original par son caractère collaboratif : on peut proposer lors d'une recherche, un lien que l'on juge pertinent, en l'ajoutant dans la liste des résultats, on peut signaler un lien inapproprié, le commenter ou le supprimer. Le moteur offre un historique de recherche et s'attache à proposer des liens de qualité, bien que la redondance d'un site dans la liste des résultats soit fréquente.

Les moteurs de recherche alternatifs

Blackle, basé sur l'index Google, propose une interface noire, qui économiserait l'énergie de notre ordinateur en watts. Des études contradictoires³ laissent à penser que cette économie d'énergie n'est réelle que par l'utilisation d'écrans cathodiques et non sur des écrans LCD. Un compteur sur l'interface de recherche affiche un compteur de Watts économisés par heure.

Doona.fr, moteur français utilise l'index Exalead. Les revenus publicitaires générés par les recherches sont reversés à 100% à des associations humanitaires.

Veosearch, moteur français, utilise l'index Yahoo. Les revenus publicitaires générés par les recherches sont reversés à 50% à des projets associatifs liés au développement durable. Il est possible de choisir les associations à qui on préfère donner en paramétrant son compte. **Ethicle**, quant à lui, plante des arbres en fonction du nombre de recherches (index Yahoo).

Les moteurs de recherches avec affichage des résultats innovant

³http://www.business-garden.com/index.php/2007/08/13/fond_noir_ne_serait_pas_ecologique prouve que l'économie d'énergie sur des écrans LCD n'est pas prouvée, alors que <http://ecoiron.blogspot.com/2007/01/black-google-would-save-3000-megawatts.html> soutient que Blackle économise de l'énergie.

Addictomatic affiche et synthétise les résultats sur une seule page sous forme de widgets, par catégories de contenus, comme dans les univers Netvibes. Chaque bloc peut être déplacé ou supprimé, pour obtenir une page personnalisée

Cuil (*Connaissance* en gaélique) moteur développé par des anciens employés de Google qui posséderait un index de 120 milliards de pages (soit trois fois plus que Google, mais invérifiable). Les résultats de recherche sont présentés sous forme de vignettes avec des résumés, classés selon des onglets thématiques. Seul problème : il n'existe qu'en anglais pour le moment, son page rank est moins efficace que Google, il ne propose pas d'autres options que la recherche Web pure, mais il respecte la vie privée car il n'enregistre pas de données de navigation. Cependant l'affichage innovant des résultats par colonnes est très critiqué par des études d'eye tracking⁴.

Oamos (uniquement en anglais et allemand). C'est un moteur de recherche fantaisiste, qui allie plusieurs gadgets à l'affichage des résultats : musique, vidéos, photos, nuages de mots-clés... Intéressant à utiliser pour des recherches concernant les loisirs, plus que pour des recherches universitaires.

MelZoo, en version Bêta, propose une interface de recherche épurée avec une barre de saisie agrandie pour un meilleur confort visuel. La recherche se fait uniquement sur le web général et est présentée dans une page séparée en deux, ce qui le rend plus ergonomique : d'un côté vous avez la liste des résultats, de l'autre la page web s'affiche en *pré-vue* pendant deux minutes maximum lorsque vous passez le curseur sur les résultats de recherche à gauche.

Twingine affiche ses résultats dans une page divisée en deux, avec d'un côté les résultats du moteur Yahoo et de l'autre l'affichage des résultats avec Google. Moteur idéal pour comparer les différences de référencement et de pertinence des résultats.

Kartoo est un métamoteur⁵ qui présente les résultats des recherches sous forme de carte heuristique. L'affichage des résultats se fait sous forme de pages vignettes reliées entre elles par liens sémantiques générés dynamiquement. Lorsque l'on survole une vignette, elle s'affiche en miniature à la gauche de l'écran. Des onglets proposent des recherches d'images et de vidéos qui basculent dans une nouvelle fenêtre sur Yahoo. En général, les résultats sont pertinents, mais l'affichage en carte est perturbant (juste l'Url et la vignette, pas de nom de site visible), mais il est appréciable de se voir proposer des mots-clés pour affiner la recherche

II. Les moteurs de recherche spécialisés

⁴ <http://motrech.blogspot.com/2008/09/trecp-de-cuil-tue-le-cuil.html>

⁵ *Un métamoteur est un moteur qui interroge plusieurs moteurs et annuaires de façon simultanée et n'a pas d'index propre, car il utilise celui des autres. Autre exemple de métamoteur : Copernic.*

Presse en ligne

[Google actualités](#) recherche les informations sur les pages web des journaux en ligne.

[Paperblog](#), [Wikio](#) et [LePost](#) : portails d'informations publiées dans la blogosphère.

Les moteurs de recherche de personnes

Les pages [jaunes/blanches](#) : recherche les coordonnées d'une personne ou d'une entreprise.

D'autres moteurs se sont développés dans la recherche d'informations sur une personne, à travers les traces laissées dans les réseaux sociaux : facebook, copainsd'avant, blogs personnels, annuaires d'entreprises... Ce sont des moteurs intéressants à connaître pour apprendre à gérer son E-réputation.

[123people](#) : site en français. Organise les résultats de la recherche en différents domaines : photos, numéros de téléphone, mails, blog, biographie... Au vu des résultats donnés, l'essai n'est pas très concluant. La plupart des photos publiées viennent des réseaux sociaux auxquels on est affilié. Le nuage de tags associé au profil rapporte des informations tout à fait mineures et peu satisfaisantes

[Whozat](#) : l'affichage des résultats se fait sous forme de liste, mais il est possible d'affiner la recherche en ciblant une ville, un âge ou le sexe, ainsi que de personnaliser la page des résultats en approuvant ou en supprimant un résultat. Les résultats sont en général très pertinents, d'autant plus qu'il est possible de rajouter des mots-clefs pour différencier les homonymes.

Moteurs de recherche de musique

[Musicme](#) et [deezer](#) sont des webradios, avec moteurs de recherche intégré, ce qui permet de rechercher et écouter en streaming les morceaux choisis. Il est possible de se créer des playlists en ouvrant un compte, d'adhérer à une communauté, de faire des commentaires et d'acheter les titres.

[Playzik.net](#), [MP3.com](#) et [AirMP3](#) sont des moteurs de recherche de fichiers MP3

[Lyrics.cd](#) recherche les paroles de chansons

[Freescores](#) recherche les partitions de musique gratuites et payantes.

Moteurs de recherche sur la santé

[Honcode](#) est un moteur de recherche de sites médicaux tels que [Caducee.net](#), [e-sante.fr](#), [creapharma...](#) Possibilité de chercher avec [Medhunt](#), intégré au moteur Honcode, qui recherche des documents médicaux. ****

Les moteurs de recherche de l'information scientifique et technique

Spinoo est un moteur spécialisé pour l'Education Nationale. Il recherche dans un index composé des pages web de 90 sites éducatifs institutionnels français (MEN, CNDP, CRDP, CDDP, INRP, Onisep, serveurs académiques...).

In extenso.org est un moteur de recherche spécialisé en sciences humaines et sociales développé par le Centre pour l'édition électronique ouverte (CLEO). Il explore une partie de la Toile mondiale, sélectionnée et indexée en un corpus réduit de sites et de dépôts d'articles par l'équipe de Revues.org. In-extenso distingue deux espaces de recherche. Le premier espace indexe des dépôts d'articles scientifiques et permet une recherche sur le titre, l'auteur et le résumé provenant de Persée, Revues.org, HAL-SHS, Gallica... (environ 135.000 pages). Il s'agit de dépôts relevant du mouvement des archives ouvertes (OAI) soutenu par le CNRS depuis la déclaration de Berlin. Le second indexe les sites sélectionnés par L'Album des sciences sociales (2 000 sites et plus d'un million de pages web).

Google Scholar est un moteur de recherche spécialisé dans les travaux universitaires : il indexe des articles issus de périodiques avec évaluation par les pairs, des thèses, des livres, des rapports, des articles en pré-publication. Avec l'accord des éditeurs, la numérisation et l'indexation de cette littérature scientifique révèle une partie du web invisible. L'index est marqué par la prépondérance de ressources en sciences exactes et médicales, par rapport aux sciences humaines et sociales. La langue majoritaire est l'anglais. On peut trouver notamment dans les sources françaises la totalité de l'archive française HAL du CCSD et les collections du fonds documentaire de l'INIST. Google Scholar propose aux bibliothèques de participer à l'alimentation du site, grâce à des liens avec leur catalogue.

OAIster est un moteur spécialisé dans le moissonnage des entrepôts OAI, c'est-à-dire des serveurs d'archives ouvertes. Le moissonneur (harvester) couvre une base de données multidisciplinaire, majoritairement en langue anglaise et récupère les métadonnées (élaborées selon la norme Dublin Core simple) associées à chaque document. Chaque notice pointe vers la ressource numérique hébergée par une institution représentant principalement des documents textuels (articles scientifiques, pré et post-publications, littérature grise) mais aussi des ressources multimédia. La base compte plus de 14 900 000 notices issues de plus de 930 institutions en janvier 2008.

Scirus est un moteur de recherche produit par Elsevier Science, spécialisé dans les documents de nature scientifique publiés sur l'internet. Scirus indexe toutes les pages de sites présents sur les serveurs reconnus et référencés par Elsevier pour leur contenu scientifique, ainsi que des documents issus de ses ressources électroniques (Science direct), de bases de données (Medline, Nasa, Beilstein), des brevets américains, de serveurs de pré-publication et Open archive. 250 millions de documents sont présents, dont 20 million de références bibliographiques, majoritairement anglais.

UREC est un annuaire de l'enseignement supérieur et de la recherche, dirigé par le CNRS.

Seek : propose une recherche générale sur le web et un onglet de recherche ciblée sur l'éducation. Moteur anglophone.

Portails de moteurs

Le centre de documentation sur internet, [Lecdi.net](#) et [Zefab](#), sont des portails de recherche documentaires qui permettent un accès rapide et simultané au contenu de différents sites retenus pour leur richesse, leur intérêt documentaire et la gratuité de l'accès à la totalité ou à une partie au moins de leur contenu.