Crible : Exalead, Dir.com, Misterbot et Mirago

Retour au sommaire de la lettre

Cette nouvelle rubrique, maintenue par Marianne Dabbadie, du magazine Veille Mag et directrice de l'innovation de la société I-KM, présentera chaque mois plusieurs outils de recherche intéressants pour vos investigations sur le Web. Ce mois-ci, ce sont Exalead, Dir.com, Misterbot et Mirago qui sont passés au crible et comparés...

Nous nous proposons aujourd'hui de faire le tour de quatre solutions européennes : il s'agit de trois moteurs français : Exalead, Misterbot le dernier-né du web francophone, et dir.com, ainsi qu'une solution d'origine britannique : Mirago. A partir du mois prochain, nous explorerons des moteurs moins connus mais qui méritent d'être testés, comme Rollyo, Releton, Misterbot, Clusty, GoshMe, Seekport, Polymeta, PreviewSeek, Numika ou autres Quintura...

Nous explorerons les fonctionnalités de ces différents moteurs avant de terminer par un test comparatif de performance des différents moteurs sur des requêtes similaires.

Exalead

Créée en 1999, Exalead est une entreprise française qui possède une filiale à Milan depuis deux ans. Son moteur de recherche est basé sur l'utilisation d'un couplage de technologies linguistiques et statistiques. Exalead est également depuis plus de trois ans, fournisseur d'AOL France pour la partie recherche d'information. Avec un index de 3 milliards de pages, Exalead se classe parmi les moteurs de recherche internationaux. Il s'agit d'un moteur dont les fonctionnalités sont si riches, que nous nous proposons de les présenter par catégorie.

Fonctionnalités de Navigation

Exalead propose des thématiques de navigation grâce à une génération automatique de mot-clés. En marge de la requête de l'utilisateur, Exalead propose une liste de thèmes associés. Le moteur propose aussi une localisation géographique des sites traitant d'un thème de recherche.

Fonctionnalités d'analyse

Exalead propose une recherche en langage naturel qui repose sur une analyse linguistique "légère" ne s'appuyant sur aucun dictionnaire. La première étape de l'analyse linguistique commence par l'élimination des mots vides. De plus, Exalead procède à une élimination automatique des doublons.
Le moteur par ailleurs propose de classer les documents retournés en fonction de leur format ; fonctionnalité fort utile, également proposée par Google, mais dans le cadre d'une recherche avancée. Parmi les fonctionnalités directement accessibles à l'internaute, on trouve aussi une pre-visualisation des pages retournées par le moteur dans la partie basse de l'écran, par un clic sur la vignette associée à chaque résultat. Par ailleurs, la présentation des résultats s'affiche en fonction des souhaits de l'internaute : avec ou sans vignette de pre-visualisation.
Le moteur possède un module de lemmatisation basée sur une analyse statistique. Exalead possède aussi, comme Google, un module de correction orthographique. Une autre fonctionnalité dérivée de la correction orthographique est la recherche phonétique approchée. La reconnaissance des langues se fait de façon automatique. Parmi ses fonctionnalités originales, Exalead propose aussi des documents audio et vidéo associés au thème de la requête.

Recherche Avancée

Les fonctionnalités de recherche avancée permettent d'activer un certain type d'algorithme, soit, de façon non exhaustive : type de pertinence, langue, localisation géographique du site, recherche phonétique, recherche sur la racine des mots. L'algorithme de recherche d'Exalead, basé par défaut sur une recherche par proximité gère également les exclusions ou encore les préférences.

Dir.com

Lancé par le groupe Iliad il y trois ans, Dir (http://fr.dir.com/) est un projet dont le groupe n'ose – à son grand tort – quasiment plus revendiquer la paternité. Pourquoi tant de frilosité ? Et bien tout simplement parce que devant le succès rencontré par le fournisseur d'accès à Internet Free, société

phare du groupe Iliad, le groupe a réorganisé ses priorités. A tel point que lorsqu'on effectue une recherche sur la chaîne de caractères "dir.com" sur le moteur de recherche interne du groupe Iliad, on n'obtient aucune réponse. Mais la recherche s'effectue pourtant à partir de ce moteur. Même le portail www.free.fr n'utilise pas le moteur produit par le groupe. Le fournisseur d'accès lui a préféré une intégration de Google en page d'accueil !
Pourtant il y a trois ans les objectifs étaient bien différents : L'éditeur n'affichait rien de moins que l'ambition de challenger Google sur les pages francophones. Actuellement, l'index mis à jour toutes les quatre semaines par le *spider* Pompos comporte environ 100 millions de pages en Français. Le moteur ne propose pas de liens sponsorisés.

Fonctionnalités
Fonctionnant sur un modèle extrêmement simple Dir.com ne tient pas compte de la casse des caractères ni de l'ordre des mots. Ses algorithmes sont basés uniquement sur l'opérateur booléen "ET". Par contre, les mots qualifiés de "mots outils" comme les déterminants ou les conjonctions sont automatiquement éliminés de la recherche. Une recherche peut être effectuée sur les pages d'un seul site de la façon suivante : "site:www.monsite.com termes_de_la_requête".
Par ailleurs, afficher un point d'exclamation devant un mot permet de l'exclure de la recherche.
Dir.com active le filtre parental par défaut. Il suffit d'aller dans les préférences pour le désactiver.
Trois masques de présentation sont actifs. Les langues actives sont le Français et l'Anglais, à choisir dans les préférences. L'une des grandes originalités de ce moteur est de proposer une option "tenir compte des accents". Dans ce cas le moteur n'affiche pas les résultats sans accents. La recherche s'effectue sur toutes sortes de formats (HTML, pdf, Word, Excel etc.).
Autre fonctionnalité originale : il est possible de formuler une requête sur dir.com en la tapant directement dans la barre d'adresses du navigateur selon la syntaxe suivante : ma_requete.dir.com
Les pages sont réputées être consultables en cache mais toutefois un appel à la fonction "cache" renvoie systématiquement le même message laconique "La page n'est pas/plus dans notre cache". Nous n'avons pas constaté d'exception.

Mirago

Lancé dans plusieurs pays d'Europe en 1997 par une entreprise Britannique, Mirago est un moteur de recherche original à plusieurs points de vue. Différents partenariats permettent à l'éditeur d'afficher des fonctionnalités basées sur les usages. Parmi les partenaires de Mirago on trouve : Lycos, AdP (Annuaire des Professionnels), des annuaires thématiques comme meilleures-offres.net, des meta-moteurs, comme Copernic ou encore Deepindex. La société Mirago est également spécialiste du marketing de mot-clés. La taille de l'index du moteur de recherche sur le web francophone est d'environ 100 millions de pages.

Quelques points importants parmi les chiffres clés fournis par l'entreprise :
- Plus d'un demi-milliard de requêtes par mois.
- Plus de 500 partenaires actifs en Europe
- 14 000 campagnes publicitaires actives

Fonctionnalités
Mirago permet d'effectuer une recherche thématique régionale grâce à une carte de France cliquable ou encore une recherche par secteur d'activité. Les secteurs proposés sont les suivants : finance, juridique, habitat, agro-alimentaire, médecine, tourisme, agriculture. Une recherche sur des annuaires thématiques peut également être effectuée. Par ailleurs Mirago procède à une "normalisation" des termes de la requête (également appelée "lemmatisation"), qui consiste à réduire les mots de la requête à leur racine.

La recherche avancée permet d'effectuer une recherche sur tous les termes de la requête, ou bien seulement sur certains mots. Le moteur gère également les exclusions (y compris sur des expressions) et la recherche en *exact match*. Une page de préférences permet de paramétrer le nombre de résultats affichés par page, d'ouvrir le site dans une nouvelle fenêtre ou encore d'activer le filtre parental qui n'est pas activé par défaut. Cependant, même en l'absence de filtre les réponses portant sur le sexe et la drogue sont automatiquement reléguées en fin de résultats et pour ainsi dire jamais affichées.

Le moteur ne tient pas compte de la casse des caractères. La page de résultats d'une recherche affiche en en-tête des liens sponsorisés.

Mirago est également accessible sur PDA à partir de l'adresse suivante : mobile.mirago.fr

Misterbot

Créé par la société SM Conseils (http://www.smconseils.net) avec 27 millions de pages indexées, Misterbot (http://www.misterbot.fr), le dernier-né du web français a la stature nécessaire pour devenir après Exalead l'un des challengers du web Français.

Fonctionnalités
L'une des originalités de ce moteur est de proposer un système d'archivage de résultats en ligne. Il est nécessaire pour cela d'avoir préalablement créé un compte. Les sites archivés seront alors consultables à partir de n'importe quel PC.

Un menu déroulant permet de sélectionner les résultats affichés selon leur extension ou leur nature : .fr .be .ca ... et mots clés dans url : blog, forum, annuaire. La soumission de sites sur Misterbot est gratuite.

Par ailleurs une fonction "listing" permet d'afficher les résultats sans le résumé et d'obtenir ainsi 30 résultats par page au lieu de 10.

Misterbot ne propose pas pour l'instant de fonctionnalités de recherche avancée.

Test comparatif

Ce test simplifié ne porte que sur les dix premiers résultats de chaque moteur de recherche à partir d'un corpus de dix requêtes. La métrique utilisée est un rappel à dix documents : les pourcentages du graphique indiquent une moyenne correspondant au nombre de documents pertinents ramenés par le moteur de recherche sur la première page de résultats. Les tests sont effectués à partir de la configuration par défaut de chaque moteur de recherche.

Les plus grandes surprises de ce test :

- Les meilleurs résultats sont pour dir.com, qui, avec 85% de bonnes réponses devance de 7 points Exalead sur ce corpus de requêtes. Ainsi l'espoir abandonné de la R&D du groupe Iliad recèlerait-il des ressources insoupçonnées ?

- Le palmarès des moins bons résultats revient au dernier-né : Misterbot, avec seulement 46 % de bonnes réponses. Surprise certes, mais pas tant que cela. Il s'agit d'une technologie en cours d'évolution, encore dans sa version bêta pour certaines fonctionnalités. La leçon à tirer de ce test est sans doute que les capacités d'indexation alliées aux statistiques ne font pas tout. Quelques algorithmes, notamment d'exclusions des mot outils, permettraient certainement de réduire le bruit.

Quelques bugs :

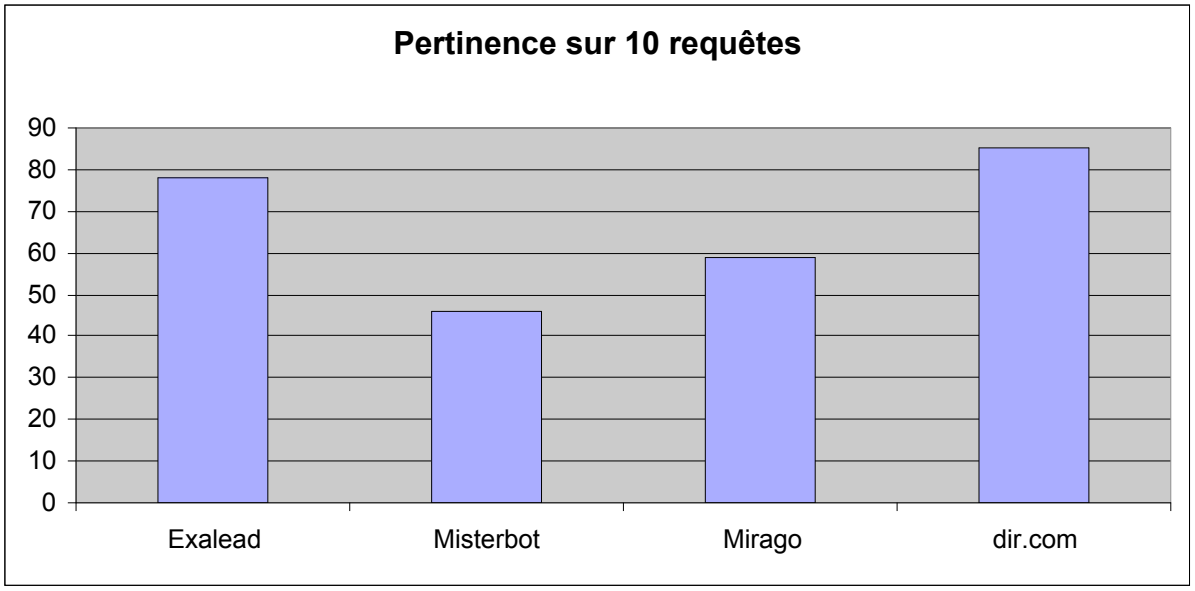
- Sur Exalead la requête "restaurants en Ile-de-France" renvoie quelques sites qui ne donnent que des adresses d'hôtels. Cela est sans doute dû à un problème lié à la proximité sémantique des deux termes.

- Sur plusieurs moteurs, la requête "Programme de cinéma" renvoie vers des chaînes de télé qui annoncent leur programme et parlent de l'actualité du cinéma. Il ne s'agit là que d'une pertinence relative que nous n'avons pas validée. Par contre sur Exalead, le menu de recherche contextuelle propose : programmes de cinéma, grilles de programme et cinéma français ; ce qui est beaucoup plus pertinent et donc plus intéressant pour l'internaute. Rappelons cependant que nos tests sont effectués sur les résultats qui utilisent la configuration par défaut du moteur de recherche et ne tiennent donc pas compte des critères de recherche affinée.

- Sur Mirago, à la requête "excursions en montagne" c'est "excursion" qui est pris en compte et seules sont proposées des excursions en mer. Sans doute cela est-il lié à une absence de

pondération des mot-clés les uns par rapport aux autres. Peut-être la recherche aurait-elle dû être effectuée en cliquant sur le département de la Savoie. Dommage pour un moteur qui présente tant de fonctionnalités originales par ailleurs.

Résultats généraux



En guise de conclusion

Au vu des tests effectués "in vivo" sur ces requêtes et sur quelques autres, on peut avancer sans grand risque de se tromper que Mirago et Dir.com qui fonctionnent sur un index francophone, affichent des résultats qui, sur le plan des usages, sont peu ou prou conformes aux attentes de l'utilisateur. Ce qui fait pourtant la différence de Mirago en termes de navigation, ce sont ses fonctionnalités originales de localisation géographique et de recherche thématique, ainsi que son accessibilité à partir des téléphones mobiles.

Il est indéniable cependant, que sur le plan de la navigation, les performances de Exalead ne sont plus à prouver et que les fonctionnalités de navigation proposées par ce moteur le classent, sur le plan des usages, parmi les meilleurs moteurs de navigation actuellement sur le marché. Nous ne pouvons cependant que signaler les performances étonnantes du moteur dir.com sur le web français, dont l'avenir, moyennant quelques efforts promotionnels, est certainement prometteur et dont l'histoire demeure sans doute encore à écrire. Le rêve initial qui consistait à challenger Google sur le web français pourrait-il devenir réalité ?

Marianne Dabadie
Directrice Innovation i-KM
Laboratoire CERSATES - UMR 8529