Ces moteurs qui clusterisent...

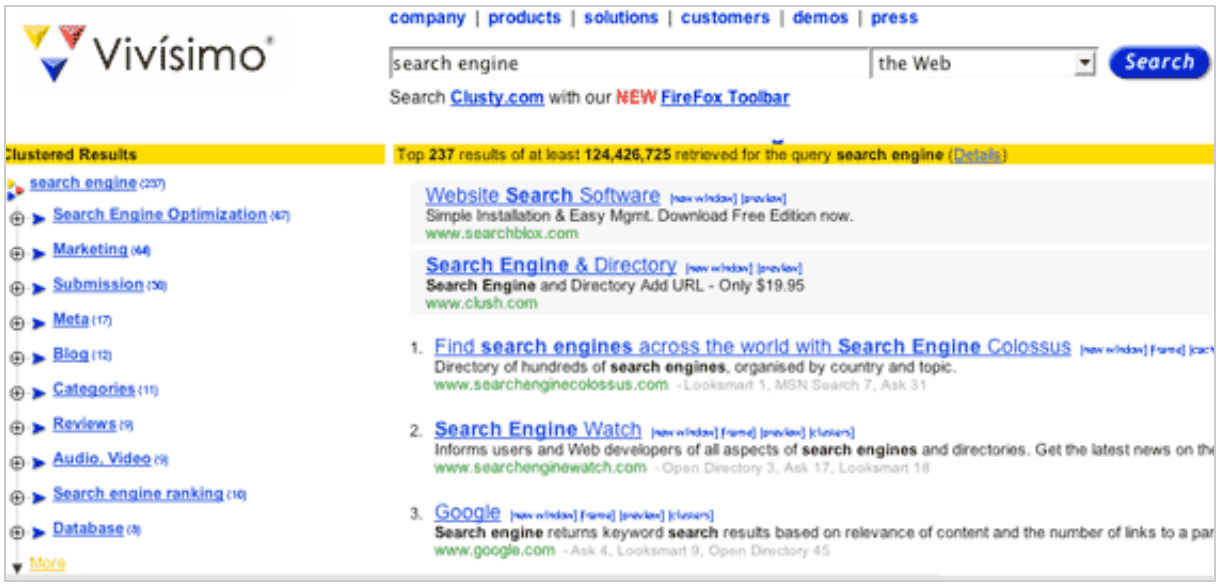
Retour au sommaire de la lettre

La plupart des moteurs de recherche majeurs travaillent sur le concept de "clusterisation", ou "découpage" des résultats en dossiers thématiques divers créés à la volée, par analyse du contenu des pages, permettant d'affiner sa recherche sur un domaine particulier. Le moteur NorthernLight, aujourd'hui quasiment disparu, a été l'un des pionniers de ces technologies, dont le flambeau a depuis été repris par des outils comme Vivisimo, Clusty, Polymeta ou Previewseek. Petite revue d'effectif...

Nous nous proposons d'explorer dans cet article trois solutions de recherche d'information : Clusty, de Vivisimo, Polymeta et Previewseek. Ces moteurs ont un point commun : ils proposent des technologies de "clusterisation". Les deux premiers sont des métamoteurs. Nous présenterons d'abord succinctement leurs fonctionnalités avant de procéder à un petit test d'utilisation sur chacun d'eux.

Vivisimo, la source de Clusty

Vivisimo (http://www.vivisimo.com/) est un produit et une entreprise ; un métamoteur de recherche et une start up, fondée en juin 2000 par des chercheurs en informatique de l'université de Canegie Mellon à Pittsburgh, université dans laquelle le moteur Lycos a également vu le jour. L'entreprise propose trois solutions, respectivement Vivísimo Clustering Engine, Vivísimo Content Integrator et Vivísimo Velocity, qui s'adressent pour le premier au grand public et pour les deux autres à des publics d'entreprise. Les principaux clients de Vivisimo sont d'ailleurs les grands comptes et administrations du "fortune 500" aux Etats-Unis.



La technologie utilisée par le métamoteur de Clusterisation Vivisimo est basée sur une nouvelle approche de clusterisation à la volée ou auto-categorisation. Le moteur définit des catégories basées sur le sens sans pre-traitement ni indexation d'une base documentaire. Cette solution de clustering n'est basée sur aucune taxonomie ni aucun thesaurus bien que capable de tirer profit de la catégorisation d'une taxonomie existante.

L'originalité de ce métamoteur réside dans sa méthode de clusterisation ou classement des sites. Vivisimo utilise uniquement les titres des pages et le résumé fourni par les moteurs interrogés pour procéder à une catégorisation effectuée sur la base d'algorithmes qui utilisent un dictionnaire de synonymes et un outil de lemmatisation. Un peu de traitement de la langue donc, allié à un solide algorithme statistique.

Des fonctionnalités basées sur les usages

En 2004, Vivisimo a lancé Clusty (http://www.clusty.com/). Clusty, c'est toute la puissance de Vivisimo alliée à la souplesse des usages. Clusty, à partir de la requête d'un utilisateur interroge le web, des blogs, des serveurs de news, effectue des recherches d'images et envoie même votre requête sur l'encyclopédie en ligne Wikipedia. Il effectue aussi une recherche sur les recherches d'emploi et sur des sites de vente en ligne. Ainsi la requête "clustering search engines" donnera selon l'onglet choisi des informations générales sur le sujet des moteurs de clusterisation, des livres à acheter en ligne sur le sujet et proposera même des offres d'emploi dans le domaine de la recherche d'information !

Dans la fenêtre des résultats, les deux premiers sont des liens sponsorisés. Sur la colonne de gauche une catégorisation est proposée. Elle est par défaut thématique mais peut être également affichée en fonction de la source ou du type d'url des résultats. Par défaut, seuls les dix premiers clusters sont présentés. Une option permet cependant de les afficher tous.

Du côté des moteurs interrogés on trouve : par défaut le web avec MSN Search, Ask Jeeves, Gigablast, Looksmart, Wisenut, ou Open Directory. Seule surprise : le métamoteur n'interroge pas Google. Une correction orthographique est effectuée grâce au Vivisimo Spelling Corrector.

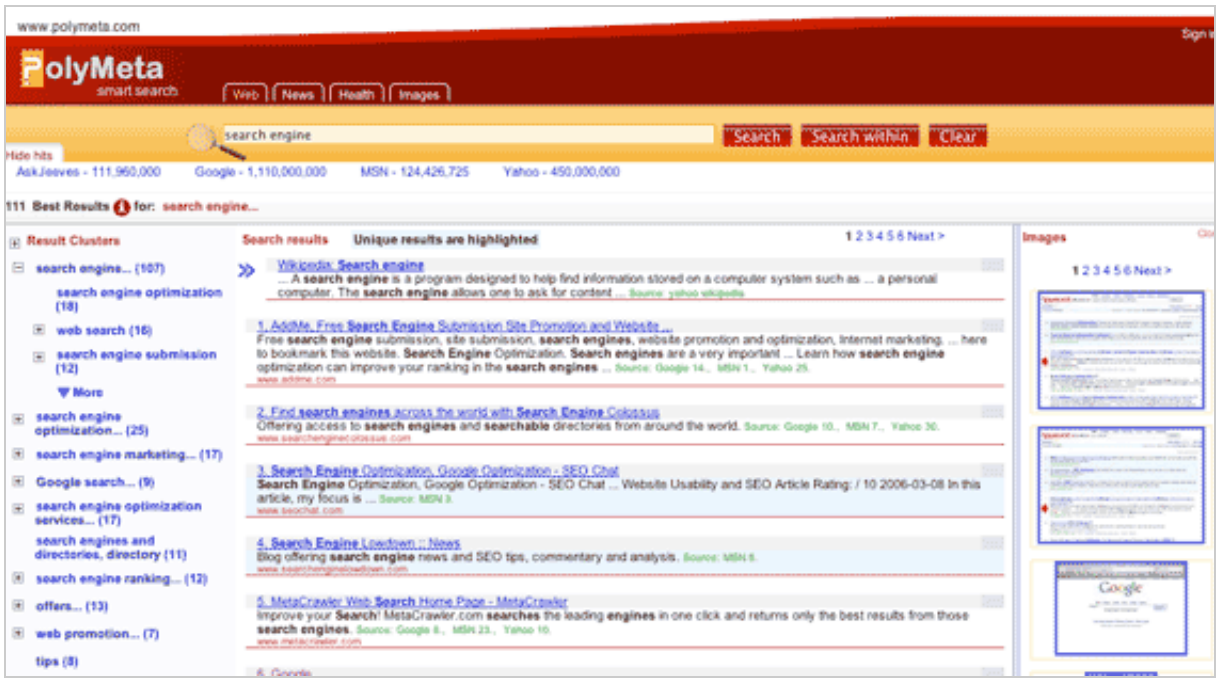
Lors de l'affichage des résultats une icône permet de previsualiser le site choisi ou de l'ouvrir dans une nouvelle fenêtre.

On trouve aussi au choix dans les préférences, des serveurs de News comme Reuters, NY Times, Yahoo News, et la possibilité d'effectuer des recherches sur des bases thématiques. Il est également possible de choisir parmi une trentaine de langues. Une option toolbar proposée en page d'accueil du moteur de recherche permet de télécharger une barre d'outils utilisable sur Internet Explorer et Firefox.

Polymeta

Polymeta est un nouveau métamoteur Hongrois. Actuellement une version en anglais est proposée en plus de la version hongroise, à l'adresse suivante :
http://www.polymeta.com/

Polymeta a été conçu par l'équipe k-prog (http://www.k-prog.com/). Sur Polymeta, la classification des sites et la clusterisation des résultats est effectuée sur la base d'une analyse contextuelle et conceptuelle. Son avantage concurrentiel sur Clusty est peut-être dans le fait que Polymeta interroge également Google. Polymeta allie la puissance des algorithmes statistiques à une analyse linguistique à trois niveaux à savoir : morphologique, syntaxique et sémantique. Programmé en Java Polymeta est portable sur toutes les plateformes.



Les fonctionnalités de Polymeta

Polymeta interroge Google, Yahoo! Search, MSN Search, AskJeeves,Gigablast ainsi que Teoma et des serveurs de News comme Google News, Yahoo!News, MSN News, NYTimes, USNews et Topix. Il propose également une recherche sur les images au travers de Yahoo, MSN et Google. Il permet aussi d'effectuer des recherches thématiques dans le domaine médical en interrogeant des bases de données spécialisées comme PubMed, MEDLINEPlus, National Institutes of Health (NIH), ClinicalTrials.gov, Scirus, Yahoo! Health, HealthMaps ou encore HealthFinder.

Tout comme Clusty, Polymeta propose les résultats de façon classique par ordre de pertinence mais également de façon thématique sur la base d'une clusterisation, avec un affichage des clusters sur la colonne de gauche, par ordre de pertinence. Le nombre de résultats par moteur est également affiché. Des images sont proposées dans la colonne de droite. Le tout a cependant tendance à surcharger un peu l'écran.

Les résultats proposés par un seul moteur sont surlignés. Une interprétation de la requête est proposée si besoin est avec reformulation et thèmes associés. Il est possible d'afficher les résultats d'un seul moteur. Un clic sur ces résultats redirige l'internaute vers le moteur les ayant proposés. Une barre d'outil Polymeta peut être intégrée à votre navigateur.

Previewseek

Previewseek (http://www.previewseek.com/) est un nouveau moteur de recherche de clusterisation, développé par une société anglaise basée à Londres, qui en propose une version beta. Contrairement à Clusty et Polymeta, il ne s'agit pas d'un métamoteur.

La technologie de Previewseek est le résultat de vingt années de recherche universitaire en Intelligence Artificielle, fouille de données, désambiguisation sémantique et utilise les résultats de la recherche en ergonomie cognitive. Tout comme Exalead qui il y a encore peu de temps s'annonçait bien plus pertinent que Google, Previewseek s'auto-proclame un peu pompeusement "the world's most advanced search engine".

Fonctionnalités : une analyse originale

Comme c'est également le cas pour Clusty et Polymeta, Previewseek propose un affichage des résultats par ordre de pertinence et sous forme de clusters dans la colonne de gauche. Jusque là rien de nouveau par rapport aux deux outils présentés précédemment. A gauche de chaque résultat une pré-visualisation de la page interrogée est proposée sous forme de vignette, comme dans Exalead. Là aussi, lors de l'affichage des résultats une icône permet de previsualiser le site choisi ou de l'ouvrir dans une nouvelle fenêtre. Un plus : on peut aussi maximiser la taille de la prévisualisation sans quitter la fenêtre de recherche.

Une originalité de ce moteur : lorsque le terme interrogé figure dans le dictionnaire, Previewseek en propose une définition qui s'affiche au-dessus de la page de résultats. Mais cette fonctionnalité semble s'appliquer uniquement aux requêtes mono-terme. Lorsqu'un mot est polysémique, le moteur en propose les différents sens possibles. Ainsi la recherche peut-elle être re-initialisée en fonction du sens souhaité. Il s'agit là d'une fonctionnalité extrêmement originale et qui correspond aux besoins d'affinage de la requête. Un bémol cependant : le dictionnaire n'est disponible qu'en anglais.

Trois onglets permettent d'effectuer une requête sur des images, de la video ou encore des fichiers MP3.

Les options de recherche avancée proposées sont des options classiques, à savoir, le type de résultat souhaité (video, web, mp3 ou images), le nombre de résultats par page. On ne trouve pas par contre le type de fichier souhaité comme c'est le cas dans Google. Sur Previewseek il est possible de sauvegarder les résultats d'une recherche sous forme de cookies.

La recherche avancée permet aussi de paramétrer le pays, donc, les sites explorés en fonction de la langue. Aucune précision cependant n'est donnée quant aux différentes langues traitées par le moteur qui ne semble réellement analyser que l'anglais…. Ce qui semble à première vue, l'éloigner quelque peu de sa performance auto-proclamée comme moteur de recherche le plus avancé du monde… Considéré, disons, dans sa globalité ! Un départ intéressant donc, mais sans doute à approfondir dans le sens du multilinguisme.

Test d'utilisation

Les différents moteurs traitant de préférence la langue de Shakespeare, notre test est effectué en anglais. Nous avons essayé de façon empirique quelques requêtes sur chacun des moteurs. Les

réponses des différents moteurs donnent une idée assez précise de leur fonctionnement. Voici ce que nous avons constaté.

Sur Clusty, les liens sponsorisés ne sont pas toujours pertinents. Dommage, car ils sont compris dans les dix premiers résultats. Un affichage des dix résultats produits par le référencement organique, permettrait pourtant de faire sortir les liens sponsorisés des calculs de performance du moteur. Sur les trois moteurs, les résultats de la clusterisation sont intéressants. Ils apportent toujours une information supplémentaire qui permet d'affiner la requête. Cependant sur Clusty, à la question Online train reservation les clusters sont plus ou moins pertinents. Parfois seul online est retenu.

Sur Polymeta, on ne trouve pas de liens sponsorisés. La question "online train reservation" renvoit 10 résultats pertinents, comme sur Clusty du reste. Sur Previewseek, le premier lien est invalide. A la question "New-York and London Stock exchange rates" Clusty identifie "New-York" d'emblée comme mot composé et il apparaît entre guillemets dans la requête. La recherche sur l'onglet "shopping" à partir de cette requête, n'est pas très pertinente. La recherche sur les news par contre est intéressante. A cette même question on trouve sur Previewseek un éclairage intéressant sur le cours du cacao. Parmi les clusters de Polymeta on trouve pour cette même requête, "Financial News" ou encore le cours du Nasdaq.

A la requête "Stomach disease research in Germany" stomach disease est associé à gastric desease sur Clusty. Par contre peu de sites parlent effectivement de recherche. Nous avons volontairement orthographié "disease" de la façon suivante : "desease". Polymeta et Previewseek ont proposé une correction orthographique. Clusty a redressé automatiquement l'erreur sur les liens sponsorisés mais pas sur les autres résultats. Sur Polymeta cette requête fait planter le serveur sur l'onglet "Health". Dommage.

Chaque moteur a ses points forts et ses faiblesses. D'une façon générale cependant la clusterisation thématique est un point fort pour tous. Elle est généralement pertinente et apporte de l'information supplémentaire.

Pour Clusty, la faible pertinence des liens sponsorisés vient affaiblir la pertinence globale des dix premiers résultats. Par contre la division thématique sous forme d'onglets est l'un des points forts de ce moteur qui interroge de multiples sources d'information.

La performance de Polymeta semble plutôt bonne et sa capacité à rivaliser avec les plus grands, réelle. Seul point négatif : l'aspect un peu surchargé de la fenêtre.

La performance de ***Previewseek*** est réelle mais ses capacités de traitement linguistique essentiellement centrées sur l'anglais sont une réelle faiblesse pour ce moteur qui veut challenger de Google en termes de performance. Il n'en est cependant qu'à sa version beta. Sa prochaine édition sera en ligne bientôt.

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