Recherche et Référencement - Numéro 29 -- Juillet-Août 2002

-- Lettre d'actualité mensuelle sur la recherche d'information et le référencement de sites Web --

Au sommaire ce mois-ci :

Le contenu de cette lettre est accessible sur la zone "Abonnés" du site Abondance, à l'adresse : http://abonnes.abondance.com/archives/acturech/0207.html

La lettre "Recherche & Référencement" paraît aux alentours du 15 de chaque mois (un seul numéro pour les mois de juillet-août)

Pour tout renseignement : © Olivier Andrieu, oa@abondance.com

---

Bonnes vacances à tous et bonne lecture !!

Fast : état des lieux et projets d'avenir

Retour au sommaire de la lettre

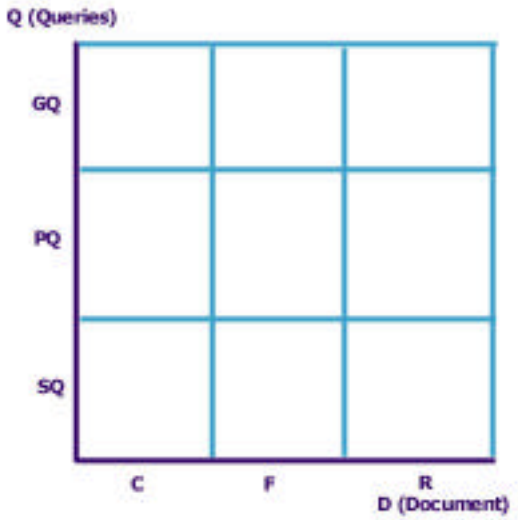
On ne présente plus Fast (http://www.fastsearch.com/), l'un des plus sérieux concurrents de Google à l'heure actuelle, pour ne pas dire l'un des seuls... D'ailleurs, son "combat" actuel avec Google pour fournir les liens "moteur" de Yahoo! est assez révélateur de la concurrence qui va se dessiner entre ces deux géants de la recherche d'information dans les mois qui viennent. Parions que l'hégémonie quasiment sans partage de Google jusqu'à maintenant sera quelque peu perturbée par un troublion nommé Fast, et son "laboratoire de recherche" AllTheWeb (http://www.alltheweb.com/). Pour en parler plus longuement, nous avons rencontré Franz Guenthner, "Professor for Computational Linguistics" à l'université de Munich, et qui est également très lié à Fast avec qui il travaille quotidiennement à améliorer l'outil de recherche.

Franz n'est d'ailleurs pas un inconnu dans le domaine des moteurs de recherche, loin de là. Il a travaillé pendant 3 ans avec l'équipe de Louis Monier, qui a développé le célèbre Altavista. En mars 1999, Louis Monier part avec son équipe (il travaille aujourd'hui pour eBay). En juin 99, Franz Guenthner feuillette négligemment un journal et tombe par hasard sur un article indiquant que Dell et Fast désirent mettre en place le plus gros moteur de recherche sur le Web mondial. Contact est pris et Franz devient rapidement "Chief Technology Advisor", terme que l'on peut traduire en "consultant extérieur" pour Fast Search and Transfert Inc., dont le siège social est basé à Oslo, Norvège. Fast a également des bureaux à Oslo, Trondheim, Munich, San Francisco, Boston, Tokyo, Rome (http://www.fastsearch.com/about/locations.asp) et un représentant à Paris (voir lettre "R&R" du mois dernier). Fast compte 200 employés dans le monde dont 50 "PHd" et 100 chercheurs/développeurs.

Pourquoi les moteurs de recherche sont-ils (encore) si mauvais ?

La vision du monde des outils de recherche par Franz Guenthner est à la fois passionnante et pertinente. Pour lui, si les moteurs de recherche sont encore "si mauvais" aujourd'hui, c'est parce qu'ils traitent de la même façon toutes les requêtes, quelles qu'elles soient. Par exemple : pour les moteurs actuels, il faut absolument que les mots demandés soient présents dans le texte de la page, ce qui est une ineptie, selon Franz. Pour les moteurs, une requête n'est qu'une suite de "chiffres", de caractères, alors qu'elle possède pourtant une structure qui reflète une sémantique qu'il est nécessaire d'analyser.

Pour obtenir une vision plus avancée du monde de la recherche d'information, il est nécessaire, selon Franz Guenthner, d'examiner les rapports possibles entre les requêtes (Q = "Queries") et le document, selon une matrice semblable à celle-ci :



Les requêtes (Queries) sont réparties en trois familles distinctes, qui représentent chacune environ 1/3 des demandes effectuées sur le Web :

- Les "General Queries" (GQ), qui représentent des demandes très génériques, très peu précises : france, tourisme, mp3, etc.

- Les "Specialized Queries" (SQ) qui, elles, sont très précises. L'internaute qui tape une telle requête sait ce qu'il veut et sait également formuler sa demande au moteur.

- Les "Problem Queries" (PQ), qui sont des requêtes émanant d'internautes sachant *a priori* ce qu'ils désirent trouver, mais ne savent pas l'exprimer. Comme leur nom l'indique, ce sont ces requêtes-là qui sont le plus complexes à traiter pour un moteur...

Côté "Contenu", il existe, là aussi, trois grandes catégories de critères de pertinence qu'il est possible de prendre en compte :

- Le Contenu de la page (C) : titre, texte, balises meta, nom des images, etc.

- Le Format de la page (F) : langue, taille (en octets), date de dernière modification, format informatique (image GIF, Jpeg, fichiers PDF, Word, Powerpoint, etc.)...

- Les Références (R) : la façon dont l'environnement Web "traite", "juge" ou "estime" la page en question : nombre de liens pointant vers elle, "qualité" de ces liens, etc.

La matrice présentée ci-dessus est donc découpée en 9 cases distinctes. Le but est alors de proposer un algorithme d'évaluation et de "ranking" différent en fonction des différentes demandes. Par exemple, une requête de type "GQ" pourra être traitée selon des algorithmes de ranking mettant en avant les références (R), car le contenu (C) n'a que peu d'intérêt, si ce n'est pour faire un premier tri dans l'index. En revanche, une requête "SP" sera certainement mieux servie par un algorithme mettant en avant le contenu (C). Etc.

Toutes ces voies sont actuellement explorées par Fast pour mettre en place une nouvelle façon d'explorer le Web et de répondre de façon pertinente aux demandes des internautes en essayant de "comprendre" la requête (soit dit en passant, bonne chance aux optimiseurs de pages web pour l'avenir ... ;-)).

Ainsi, il peut se passer deux cas distincts :

- Soit la requête est dite "atomique", c'est-à-dire qu'on ne peut pas la décomposer. Elle est proposée sous la forme d'un terme unique. Dans ce cas-là, il y a de fortes chances qu'il s'agisse d'une "General Query". Elle sera donc traitée dans cette optique.

- Soit elle peut être décomposée, et dans ce cas, il est possible de définir dans la requête deux champs différents :

* Le "Head" (H), qui est la partie (mot ou expression) qui représente l'essence même de la demande.

* Le "Content" (Co), qui représente, quant à lui, les informations qui viennent compléter le "Head" pour affiner la demande.

Exemple pour la requête : "Britney Spears Pictures" : "Britney Spears" représente le "Head", partie centrale de la demande, et "pictures" le "Content", qui se réfère au "Head".

Idem pour "Weather in Strasbourg" : "Strasbourg" est le "H", "weather" le "Co". "in" est un mot "vide" dont il ne faut pas tenir compte.

Le moteur de recherche ne devra donc pas traiter le "H" de la même façon qu'il traite le "Co". Car il ne s'agit pas de proposer à l'internaute un site sur la météo américaine, mais bien un site qui traite de météo mais obligatoirement sur la ville de Strasbourg, pour être pertinent.

Chaque requête non-atomique peut donc être décomposée en "H" et "Co". Elle peut contenir plusieurs "H", plusieurs "Co" ou un mélange des deux, bien sûr. Il s'agit là des principes fondamentaux des recherches de Franz Guenthner, qui sont peu à peu incluses notamment dans le moteur Fast et les technologies qui y sont rattachées. De cette décomposition des requêtes en "H" et en "Co" dépend une meilleure compréhension des demandes des internautes et, donc, la possibilité de placer la requête dans la matrice indiquée ci-dessus et la prise en compte d'un algorithme de ranking différent.

D'ailleurs, certains points de cette technologie sont déjà en ligne sur le moteur AllTheWeb. Si vous tapez la requête "britney spears pictures" (sans les guillemets, bien sûr) sur ce moteur : http://www.alltheweb.com/search?cat=web&cs=utf-8&l=any&q=britney+spears+pictures

Regardez bien la case "About Your Query" à droite de l'écran. Elle indique :

Your query was rewritten into:
"britney spears" pictures

Le moteur a, de lui-même, ajouté des guillemets à l'expression "britney spears" car il l'a reconnu comme étant un "H". De même, la requête "police department new york city" : http://www.alltheweb.com/search?cat=web&cs=utf-8&l=any&q=police+department+new+york+city

Fournit les infos suivantes :

Your query was rewritten into:
"police department" "new york city"

Les expressions "police department" et "new york city", qui sont manifestement 2 "H", ont été automatiquement détectées.

De la même façon, si l'on tape la requête "Where can I find information about Fast ?" : http://www.alltheweb.com/search?cat=web&cs=utf-8&l=any&q=where+can+i+find+information+about+Fast+%3F
On obtient :

Your query was rewritten into:
fast

En règle générale, regardez bien cette case "About your query" sur AllTheWeb (je l'ai moi-même découverte à cette occasion), elle est très intéressante pour en savoir sur la façon dont "réfléchissent" AllTheWeb et Fast...

Projets en cours

L'équipe de Franz Guenthner est également en train de travailler sur une catégorisation des "Co" et des "H" (plus de 4 millions de ces derniers déjà identifiés) sur la base des requêtes déjà effectuées sur Fast depuis sa création. Il sera ainsi plus facile à un moteur d'identifier les uns et les autres.

D'autre part, l'une des raisons pour lesquelles les recherche échouent sur une requête est également le problème des fautes d'orthographes. Le moteur doit donc avoir à sa disposition un correcteur d'orthographe sur les mots dans de nombreuses langues. Par exemple, le mot "Lufthansa" a été demandé sur Fast sous 1200 formes différentes. On connaît également la page de Google sur laquelle il propose toutes les orthographes demandées sur le moteur pour le nom de la chanteuse "Britney Spears" : http://www.google.com/jobs/britney.html

Mais le correcteur orthographique doit également savoir traiter les URLs. En effet, selon les études effectuées par Franz Guenthner, 2% des requêtes tapées sur le web sont des urls (du type www.hotmail.com, fnac.net, etc.) et elles contiennent parfois des fautes de frappes ou d'orthographes. Il faut donc pouvoir les corriger. Un tel correcteur sera prochainement mis en ligne sur AllTheWeb. Plus que de la correction d'orthographe "simple", c'est plutôt de "lemmatisation" (recherche et identification de la racine des mots demandés) qu'il faut parler d'ailleurs. L'outil sera donc capable de détecter les pluriels, les féminins, etc.

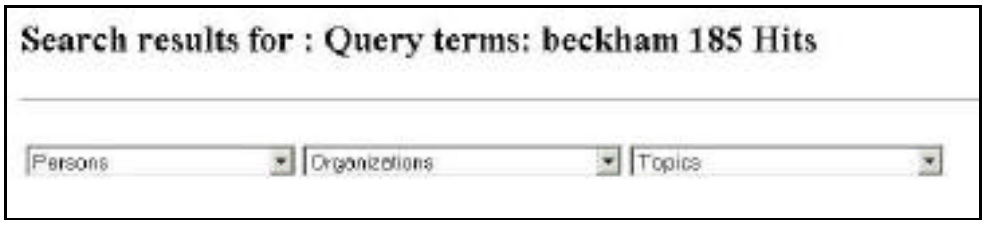
L'équipe de Franz Guenthner travaille également sur une nouvelle forme de "Related Searches". On connaît bien le principe de base des "Related Searches" : pour un mot clé générique (assurance), le moteur va proposer d'autre expressions contenant le mot demandé, sur la base des requêtes précédemment tapées par les internautes. Ainsi, Voila, sur le mot clé "assurance" (http://search.ke.voila.fr/S/voila?kw=assurance&dt=*) va proposer les expressions suivantes : "assurance auto", "assurance moto", "assurance automobile". Ce schéma, basé sur les requêtes identifiées lors des semaines précédentes, est aujourd'hui assez classique et proposé par la majeure partie des outils de recherche.

Certains essaient cependant d'aller plus loin : Exalead, avec une technologie basée sur des algorithmes statistiques (voir les numéros précédents de cette lettre), Altavista avec son système "Prima" (ex-Paraphrase), mis en ligne officiellement depuis quelques jours, Kartoo (le système n'est pas encore en ligne au moment où cet article est écrit, mais le sera dans les jours qui viennent), etc.

Le but est de ne plus donner des expressions contenant obligatoirement le terme demandé ("assurance" donne "assurance auto") mais bien des termes connexes ("assurance" donne "devis", "dommages", "contrat", etc.). Pour cela, Franz Guenthner a fait une démonstration d'un applicatif développé pour le moteur de "news" (dépêches d'actualité) d'AllTheWeb et qui est assez impressionnant : suite à une requête, il propose 3 menus déroulants : "Persons", "Organizations" et "Topics" (un quatrième, "Localisation" devrait apparaître prochainement). Pour chaque requête, le contenu de chaque menu déroulant va faire apparaître des termes pertinents. Par exemple "french politics" proposera, dans son menu intitulé "Persons", les noms de Jacques Chirac, Lionel Jospin, etc. Chaque contenu est, bien entendu, évalué et construit en temps réel. Impressionnant... L'outil devrait donc être prochainement présent sur AllTheWeb, à côté du lien "More news", proposé sur les dépêches d'actualité, en début de liste-réponse (tentez une requête sur "Bush", par exemple).

Exemple : une recherche sur le mot "beckham", qui donne sur l'outil de test de Franz, 185 résultats :

La page de réponse propose trois menus déroulants :



Le premier menu va proposer des noms de personnes ayant un rapport avec le mot clé demandé :

Le deuxième menu propose des organismes :

Le troisième des thèmes de recherche :

D'autre part, un moteur de recherche doit également être capable de fournir des réponses dans une thématique donnée. Si une requête a trait à l'informatique (exemple : "software windows"), le moteur ne doit pas proposer des liens vers des fabricants de fenêtres ("windows"). Pour cela, il est nécessaire de catégoriser de façon automatique les pages du Web. Là aussi, l'équipe de Franz Guenthner a développé un classifieur de pages assez performant, dont la démo m'a été faite sur des dépêche d'actualité de l'agence Reuters. Tout d'abord, Reuters a fourni au préalable de nombreuses catégories prédéfinies. Le système prend ensuite en entrée un texte de dépêche d'actualité, quel qu'il soit. Puis, il l'analyse et fournit en sortie les catégories dans lesquelles la dépêche doit être intégrée : politique française, sport international, etc. L'outil, dans sa première version, trouvait déjà 85% des thématiques qui, auparavant, étaient définies par des documentalistes.
Ce type d'applicatif est également très utile pour classifier des pages web dans un environnement plus ou moins homogène, notamment sur un intranet.
Exemple : le site http://www.scirus.com/ et sa recherche avancée (http://www.scirus.com/search_simple_boolean/), qui a été créé par l'équipe de Franz, et qui traite de façon spécifique de l'information scientifique disponible en ligne. Sur ce site, toutes les classifications proposées dans la recherche avancée ont été mises en place de façon automatique.

Franz Guenthner m'a également montré un outil assez impressionnant qui est un générateur d'annuaire. Sur la base de quelques informations de départs (mots clés et/ou liens), l'outil est capable de recréer une arborescence très complète de catégories et sous-catégories. Une première application a été lancée sur le domaine de la physique et le résultat était plutôt remarquable. De plus, un autre outil permet, en proposant de 1 à 3 urls d'un domaine donné, d'explorer le Web et de fournir plusieurs dizaines, voire centaines, de sites traitant de la même thématique, classés par ordre de pertinence et de popularité. Et le tout en quelques minutes, en utilisant la topologie des liens du Web (notamment l'analyse des pages qui pointent vers plusieurs sites différents). Un outil extrêmement intéressant pour les documentalistes des annuaires, notamment, qui peuvent ainsi avoir sous les yeux le "must" du Web sur une thématique précise , et être sûr de ne pas oublier un site essentiel pour la rubrique en question. Quand on sait que l'équipe de Franz travaille également sur des algorithmes qui permettent de faire ressortir les nouveaux sites d'un domaine, ceux qui ont

été mis en ligne depuis peu, on imagine facilement l'intérêt de ce type d'outil...

Fast aujourd'hui

Les projets décrits précédemments sont aujourd'hui en cours de développement et seront intégrés dans les produits de Fast (qu'il s'agisse des moteurs Web ou des outils développés sur mesure pour des clients comme IBM ou eBay) d'ici à la fin de l'année, normalement. Mais l'outil Fast, sous sa forme actuelle, est déjà présent non seulement sur AllTheWeb, mais également sur des sites comme Lycos, T-Online, Freeserve ou Tiscali, entre autres.

Point important : il n'existe qu'un seul index "Web" chez Fast. Il est situé géographiquement à Sacramento, aux Etats-Unis. Il tourne sur 700 PC (nombre que Franz Guenthner compare aux 15 000 machines de Google). Cependant, il est prévu un deuxième "Data Center" au cas où Fast signe avec Yahoo!, car l'histoire montre que les mises en place des précédentes solutions "moteur" sur Yahoo! n'a pas toujours été simple, avec quelques "crashs" mémorables par le passé...

Pour l'instant donc, tous les clients "Web" de Fast, et notamment tous les sites de Lycos, se "nourissent" du même index, même s'il est possible, pour chaque client, de "jouer avec les manettes" et de modifier les algorithmes de ranking en fonction de telle ou telle demande, ce qui explique les réusltats parfois différents d'un site à l'autre.

Attention également : sur tous les sites clients de Fast et AllTheWeb également, les algorithmes de ranking bougent énormément, car ils sont l'objet de très nombreux "réglages" et expérimentations pour obtenir des résultats toujours plus pertinents. La vérité d'un jour n'est donc pas nécessairement celle du lendemain...

Cependant, les critères de pertinence les plus importants sur Fast sont les suivants :

- L'indice de popularité pour les "General Queries". Il est calculé selon un mode proche de celui de Google : Le nombre de liens compte, bien sûr mais c'est surtout leur "qualité" qui est importante (voir notre article sur le sujet dans la lettre "R&R" du mois de février 2002 : "Comment est calculé l'Indice de popularité sur les moteurs de recherche").
- Le titre des pages
- Le contenu textuel
- La position du terme demandé dans le document : haut de page / bas de page ?
- Indice de densité du terme dans le document et sur le web. Fast maintient aussi un index special pour tous les "trigrams" dans les document pour traiter les phrases d'une façon efficace. Cela signifie qu'en addition à un index qui indique pour chaque mot ou il apparaît, et combien de fois il apparaît dans tout le web, il existe aussi un index de tous les triplets de mots (suite de trois mots) pour rapidement trouver des phrases quand les requêtes contient des "guillemets".

Enfin, le délai de mise à jour annoncé pour l'index de Fast (entre 9 et 12 jours) serait actuellement tenu.

Différences entre Fast et Google

Google et Fast sont les deux plus importants challengers dans le domaine de la recherche d'information aujourd'hui. Leurs technologies sont très proches, mais elles diffèrent légèrement, sur certains points :

- Google donne un plus fort poids à la proximité des mots lorsqu'une requête n'est pas "atomique". Sur la requête "concert madonna", une page contenant ces deux mots proches l'un de l'autre sera mieux classée sur Google. Fast exploite moins cette particularité aujourd'hui mais la proposera d'ici à la fin de l'année.

- Google exploite le critère de l'indice de popularité de façon importante sur tout son index. Fast prend en compte ce critère sur quelques millions de pages uniquement (notamment pour les "General Queries"). Là aussi, cela devrait changer d'ici à la fin de l'année.

- Google indexe uniquement les 100 premiers kilo-octets des documents Web. Fast indexe la

totalité des pages.

- Google donne une forte importance à l'"anchor text", c'est-à-dire le texte du lien qui pointe vers la page à classer (voir article de février 2002). Si une page A contient un lien vers une page B et que le lien textuel de la page A contient le mot "assurance", la page B sera bien classée sur ce terme. Fast exploite moins ce critère, mais travaille dessus.

- Fast semble plus complet au niveau de ses "sous-index" européens que Google. Selon Franz, Fast proposerait environ 20% de pages en français de plus que Google, sans pour autant indiquer leur nombre total.

- Fast et Google proposent chacun un index de plus de 2 milliards de pages, mais Fast devrait atteindre les 3 milliards d'ici à la fin de l'année, voire même avant. Mais Google travaillerait également dans ce sens...

Quelques chiffres

Quelques chiffres donnés par Franz Guenthner lors de notre entretien, et qui me semble intéressants :

- En l'an 2000, la zone de recouvrement entre les 3 principaux annuaires anglophones majeurs (Yahoo.com, Looksmart.com et Dmoz.org) n'était que de 5%. En d'autres termes, seuls 5% des sites de chaque annuaire se retrouvait donc chez les autres concurrents. Etonnant, non ?

- Lorsque Altavista avait mis en place ses "Related Searches", du jour au lendemain, 25% des internautes les utilisaient !

- La taille du Web statique serait aujourd'hui comprise entre 4 et 5 milliards de pages.

- 22% des requêtes sur le Web ont trait au sexe.

- Aujourd'hui, seuls 10% des sites du Web sont "pointés" (disposent de liens vers eux) par d'autres sites. Ce chiffre était encore de 30% il y a quelques années.

- Sur 750 millions de requêtes qui ont été étudiées par l'équipe de Franz Guenthner, il est possible d'identifier 250 millions de requêtes différentes. D'autre part, parmi celles-ci, 1 million seulement correspondent à 50% des demandes.

Le but pour Franz est donc d'essayer, avec Fast et AllTheweb, d'être le plus pertinent possible sur les 3 millions de requêtes les plus fréquentes, ce qui représente environ 2/3 des demandes des internautes.

Il s'agit là du but avoué de Fast dans les mois qui viennent. L'avenir nous dira donc s'il y est parvenu. prochaine étape de son développement : l'éachéance du contrat qui lie Google à Yahoo!. Ce contrat a été prolongé jusqu'en septembre, date à laquelle Yahoo! prendra sa décision. Google ou Fast ? Fast ou Google ? Il n'a pas certainement échappé à Google, qu'à chaque fois, le partenaire "moteur" de Yahoo! a profité de son partenariat avec le célèbre portail pour exploser au niveau de sa notoriété : Altavista, puis Inktomi, puis Google. Chacun des partenaires de Yahoo! a sans conteste bâti son succès sur le ciment de son partenariat avec Yahoo! Il peut donc être intéressant, voire nécessaire, pour Google de renouveller son contrat avec Yahoo!... Pour que Fast ne le fasse pas, un tel accord pouvant placer ce dernier en position de concurrent direct en quelques mois... Le contrat est bien sûr d'importance pour Fast. Réponse à la rentrée !

Et merci à Franz Guenthner pour son accueil et les informations fournies lors de notre entretien !!

Les applicatifs API de Google

Retour au sommaire de la lettre

Google propose depuis quelques mois le service "Google Web API" (http://www.google.com/apis/). L'initiative est intéressante et assez innovante dans le domaine des moteurs de recherche : il s'agit de mettre à la disposition des développeurs informatiques une bibliothèque de fonctionnalités (API signifie "Application Programming Interface") utilisables, comme un "Web Service", et qui permettent d'utiliser l'index de Google pour bâtir de nouveaux programmes autour de la recherche d'information sur le Web. Il est techniquement possible d'écrire des applicatifs Java, Perl, PHP, C++ ou grâce à n'importe quel langage qui supporte les "Web Services". Pour savoir ce que sont ces "Web Services", voici un dossier du Journal du Net qui devrait vous en dire plus : http://solutions.journaldunet.com/dossiers/webservices/sommaire.shtml

Principe du Google Web API

Le principe est ultra-simple : vous téléchargez tout d'abord un kit de développement, puis vous demandez une license (gratuite) en vous inscrivant auprès de Google. Cette license vous permettra d'effectuer, grâce au kit, jusqu'à 1 000 requêtes sur l'index de Google par l'intermédiaire des applicatifs que vous avez développés. Un forum de discussion (http://groups.google.com/groups?group=google.public.web-apis) a même été mis en ligne par Google pour discuter de cette nouvelle fonctionnalité spécifiquement "taillée" pour les programmeurs.

Le "Kit de développement" contient des fichiers d'aides, une bibliothèque de fonctions, des exemples, etc. Bref, tout le nécessaire pour laisser vagabonder sa créativité informatique...

Condition *sine qua non* : les applicatifs créés ne doivent pas être à objectif commercial sans l'autorisation écrite de Google. Ceci dit, la restriction à 1 000 requêtes par jour est assez limitative dans ce sens. Autre limite : il n'est pas possible d'accéder à plus de 10 liens par requête (en d'autres termes, on ne peut afficher 50 résultats pour un mot clé), et il n'est pas possible non plus d'aller au-delà du 1 000ème lien issu de l'index pour une requête donnée. La requête, de plus, ne doit pas contenir plus de 10 mots et faire plus de 2048 caractères, ce qui laisse quand même une certaine marge...
Toujours au niveau des restrictions, notons également que les API ne donnent accès qu'à l'index "pages web" de Google (2 milliards de pages), pas aux forums de discussion (Google Groups), ni à l'Open Directory (Google Directory) ou aux images (Google Image Search). De plus, il est interdit d'utiliser ce service pour créer un concurrent à Google. Si, si, c'est écrit dans les conditions d'utilisations (http://www.google.com/apis/api_terms.html) !!

Les API nous en disent plus sur la syntaxe d'interrogation du moteur

Les bibliothèques de fonctionnalités fournies aux programmeurs sont assez complètes. Point très intéressant, les documents qui les décrivent nous permettent d'ailleurs d'en savoir plus sur la syntaxe d'interrogation du moteur :

* Un "ET" (qu'il est possible d'indiquer par le signe "+") est utilisé par défaut, ce qui ne surprendra personne, j'imagine (l'information est bien connue)... La requête **moteur recherche** et donc équivalente à **+moteur +recherche**.

* Google ignore les "stop words", mots anglais trop souvent utilisés comme "where", "to" ou "how". Si l'on veut effectuer une requête contenant de façon expresse de tels mots, il faut utiliser les guillemets : **"to be or not to be"**. Dans ce cas, les termes **to**, **not** ou **or** seront pris en compte, ce qui n'aurait pas été le cas sans les guillemets.

* Tous les signes non-alphanumériques de la requête sont traités comme des séparateurs, à l'exception du signe +, du signe -, des guillemets (") et de l'esperluète (&).

* La syntaxe d'interrogation de Google est la suivante :

ET = signe + : **Star Wars Episode +I.** A priori inutile employé seul, puisque que le ET est l'opérateur par défaut de Google.

SAUF = signe - : **basse -musique**

Expressions = guillemets (**"moteur de recherche"**)

OU = Booléen OR : **vacances londres OR paris** : **vacances** est obligatoire ici, le OR s'applique aux deux termes suivants (Londres OU Paris)

Restriction sur un site : **google site:www.abondance.com** recherche toutes les pages contenant le mot "Google" sur le site www.abondance.com. Autre possibilité avec le signe - : **google -site:www.google.com**. Une seule fonction "site:" est acceptée par requête. Il est également possible, depuis peu, d'effectuer des recherches sur un nom de domaine : **google site:abondance.com** propose ainsi les sous-domaines du site : actu.abondance.com, outils.abondance.com, etc. De même, **google site:org** ne donnera comme résultat que des pages extraits de sites en .org.

Restriction sur la date : **"le grand bleu" daterange:2452122-2452234**. Le premier paramètre est la date de début de la recherche, le second la date de fin de recherche. Le tout exprimé en jours juliens, donc le nombre de jours depuis le 1er janvier de l'an 4713 avant Jesus-Christ. Eh oui. Rien que ça. Si vous voulez en savoir plus sur les dates juliennes :
http://www.altcal.com/jourjul.html
http://astrojacky.logidac.com/cahier/chap01-html/x87.htm
Heureusement, il existe de nombreux convertisseurs entre les deux formats (julien et actuel). Voici trois liens intéressants :
http://www-aviso.cnes.fr:8090/HTML/information/missions/topex/outil_jjtocd_fr.html
http://aa.usno.navy.mil/data/docs/JulianDate.html
http://www.tesre.bo.cnr.it/~mauro/JD/
ou, plus simplement, ici, avec un applicatif qui permet d'effectuer des recherches sur Google en proposant des dates "normales" ;-), converties en dates juliennes "à la volée" :
http://www.faganfinder.com/engines/google.shtml

Recherche sur le titre des documents : **intitle:bourse glossaire**. Dans ce cas, le mot clé "glossaire" sera recherché dans la page en entier, et "bourse" uniquement dans le titre (attention : pas d'espace entre "intitle:" et "bourse").
Autre possibilité avec la fonction allintitle : **allintitle: bourse glossaire**. Les deux termes demandés ("glossaire" et "bourse" ne seront recherchés que dans le ltitre (attention : un espace est nécessaire entre "allintitle:" et "bourse").

Recherche dans l'url des documents : **inurl:logiciels word**. Dans ce cas, le mot clé "word" sera recherché dans la page en entier, et "logiciels" uniquement dans l'adresse de la page. ne fonctionne qu'avec le premier mot qui suit "inurl:" et n'accepte pas les ponctuations (attention : pas d'espace entre "inurl:" et "logiciels"). Il est possible d'utiliser plusieurs fonctions "inurl:" dans une requête.
De même que pour la recherche dans les titres, il existe une fonction allinurl : **allinurl: logiciels word**. Les deux termes demandés ("logiciels" et "word") ne seront recherchés que dans l'url (attention : un espace est nécessaire entre "allinurl:" et "logiciels").
Attention également : certains signes et caractères peuvent être pris pour des séparateurs. Ainsi :
allinurl: logiciels/word
est interprété ainsi :
allinurl: logiciels word

Recherche sur le texte de la page (zone "Body" du code HTML sauf le texte des liens) : **allintext: google france**. Fonctionne de la même façon que allintitle: ou allinurl:

Recherche sur les liens (texte des liens) : **allinlinks: google** effectuera une recherche sur les pages qui contiennent un lien dont le texte contient le terme "Google" (exemple : cliquez ici pour aller sur le site de Google). Fonctionne avec les mêmes restrictions que allintitle: ou allinurl:

Recherche sur le type de fichier : **confidentiel filetype:pdf** recherche les fichiers au format

PDF qui contiennent le mot "confidentiel". Le paramètre de la fonction "filetype:" est égal à l'extension du fichier en question : swf (Flash), doc (Word), rtf (Rich Text Format), pps (Powerpoint), etc.

Recherche sur un seul lien : **info:www.abondance.com** retourne un seul lien du site demandé au lieu de 2 par défaut. Ne fonctionne pas avec un mot clé fourni en plus (de type **google info:www.abondance.com**).

Recherche de l'indice de popularité : la fonction link: (**link:www.abondance.com**) permet d'obtenir toutes les pages qui ont mis en place un lien vers votre site. Si vous voulez en exlcure les pages internes du site en question, il faut utiliser la requête : **link:www.abondance.com - site:www.abondance.com.** La fonction link: semble cependant ne pas fonctionner tout le temps sur Google...

Recherche des sites similaires : **related:www.google.com** propose les sites de Yahoo!, Altavista, Lycos, etc.

Recherche dans le cache : **cache:www.abondance.com** vous propose la version du site web indiqué qui est dans le cache de Google (le version telle qu'elle a été indexée par Google). Cette fonction marche également avec des pages internes, comme dans la requête : **cache:www.abondance.com/docs/oa.html**

On s'aperçoit ici de la richesse de la syntaxe d'interrogation de Google, pourtant peu utilisée, c'est certain... Un regret cependant : aucune fonctionnalité de permet d'obtenir la valeur du "PageRank" (mesure de la pertinence chez Google) d'un site. Dommage mais compréhensible... ;-)

Les Google Web APIs permettent également d'effectuer un filtre sur les langues, les pays ou les contenus (sites gouvernementaux US, sites traitant de Linux, de Macintosh ou de FreeBSD), ainsi que la possibilité d'effectuer du clustering (deux liens seulement par site). Il propose de plus des fonctions de dédoublonnage (une seule page affichée si d'autres documents très similaires dans leur contenu sont sélectionnés). Des fonctions d'encodage de caractères (selon les pays pris en compte) sont également fournies. Enfin, un filtre familial peut être activé à la demande.

Quelques exemples de services déjà en ligne

Sur cette base, de nombreux services ont déjà été créés par des développeurs dans le monde entier. En voici quelques exemples :

Googlematic : applicatif qui permet d'effectuer des recherches sur Google à partir d'un outil "Messenger" (messagerie instantanée) AOL ou MSN. En cours d'amélioration au moment où ces lignes étaient écrites.
http://interconnected.org/googlematic/

Windows XP Smart Tags application : permet de surligner un mot dans un applicatif Micorosoft (Word, Excel...) et de lancer une requête sur Google avec ce terme comme requête.
http://www.perfectxml.com/SmartTagsGoogle.htm

GoogleMail : possibilité d'envoyer des requêtes et de recevoir des résultats de Google par mail.
http://capescience.capeclear.com/google/

Googlebox for ASP : permet de cliquer sur un lien et d'obtenir les résultats de Google dans une fenêtre pop-up sous un format très simple.
http://www.edazzle.net/default.asp#googlebox

Google Outline Browser : permet de rechercher des sites simialires.
http://www.kasei.com/google/browse
http://radio.outliners.com/googleOutlineBrowser
http://www.staggernation.com/garbo/

Proximité : recherche les pages contenant des mots clés proches les uns des autres (fenêtre de proximité programmablede 1 à 3 termes)

http://www.staggernation.com/cgi-bin/gaps.cgi

Google Smackdown : fournit le nombre de réponses proposé par Google pour deux mots clés différents.
http://www.onfocus.com/googlesmack/down.asp

GAWSH : retourne uniquement les noms de domaine des liens proposés par Google. Le choix d'un site (en cliquant sur le triangle à gauche de l'url) propose alors toutes les pages indexées par Google pour ce dernier :
http://www.staggernation.com/gawsh/

Moteur de recherche : intégration d'une interrogation Google depuis un script PHP :
http://www.xhtml.net/cote_serveur/php/php_api_google/

Moteur de recherche : comment intégrer les résultats de Google dans votre propre charte graphique :
http://www.phpinfo.net/?p=google
http://www.phpinfo.net/google/api-google.php

Cartographie des résultats de Google avec Mapstan et Touchgraph :
http://search.mapstan.net/
http://www.touchgraph.com/TGGoogleBrowser.html

Recherches quotidiennes : cet applicatif effectue des recherches sur le mot clé demandé, mais uniquement sur les pages datant d'aujourd'hui, d'hier, des 7 ou 30 derniers jours. La recherche s'effectue sur Google ou Altavista :
http://www.researchbuzz.com/toolbox/goofresh.shtml
http://www.researchbuzz.com/toolbox/altafresh.shtml
Si ces deux applicatifs ne sont pas bâtis sur la base des "Google API", il m'a semblé intéressant de les indiquer ici...

Voici également un article sur la façon de lier les Google Web API avec PHP (indisponible aux moments de nos tests) :
http://toys.incutio.com/php/php-google-web-api.htm l

Et enfin, quelques autres applications (surtout pour développeurs) :
http://www.soapware.org/directory/4/services/googleApi/implementations

Conclusion

Si le "kit de développement" de Google est intéressant et certaines applications déjà disponibles dignes d'intérêt, on reste quand même un peu sur sa faim, un peu déçu des fonctionnalités proposées et développées jusqu'à maintenant sur la base de ces Google API. A part les possibilités d'interroger Google par mail et d'intégrer les résultats du moteur dans sa propre charte graphique, voire le système de proximité ou l'applicatif de Mapstan en France, rien de très extraordinaire, il faut bien l'avouer...

Il semble évident qu'il y a certainement mieux - ou plus - à faire avec les Google Web API. Allons, les dévelopeurs, ressaissez-vous, soyez créatifs et proposez de nouvelles fonctions innovantes, car pour l'instant, tout cela semble un peu pauvre ! Point étonnant, voire significatif, d'ailleurs : il n'existe aucun site (même pas de la part de Google) qui recense de la façon la plus exhaustive possible tous les applicatifs développés. Ca intéresse quelqu'un ?

PS : Malgré de nombreuses recherches, certains programmes dignes d'intérêt développés sur la base des Google API nous ont peut-être échappés. Si vous en connaissez qui ne sont pas indiqués dans cet article, merci de nous les signaler !! Et n'hésitez pas non plus à nous envoyer des suggestions d'applications à créer, je les proposerai dans la lettre "R&R" ces prochains mois !

Voila / Wanadoo : mise en place de la V2

Retour au sommaire de la lettre

Le site Wanadoo a dernièrement mis en place la V2 de son site web (http://www.wanadoo.fr/), le basculement ayant eu lieu au cours du dernier week-end du mois de juin 2002. Voila ayant mis en place une nouvelle version de sa page de résultats il y a quelques semaines de cela, nous avons trouvé là une bonne occasion pour faire un point sur ces deux sites et leurs (nouvelles) fonctions de recherche.

Wanadoo : la V2 est en ligne !

Le site Wanadoo a donc profité de la mise en place de son nouveau logo pour relooker son site web et ses fonctionnalités de recherche. On peut, tout d'abord, noter que les interfaces de recherche de Voila (http://www.voila.fr/) et de Wanadoo (http://www.wanadoo.fr/) fonctionnent aujourd'hui exactement sur les mêmes bases de données : les versions du Guide (annuaire) et du moteur sont strictement identiques pour les deux sites.

Sur la page d'accueil du site Wanadoo, le formulaire de recherche est maintenant en haut à droite de la page :

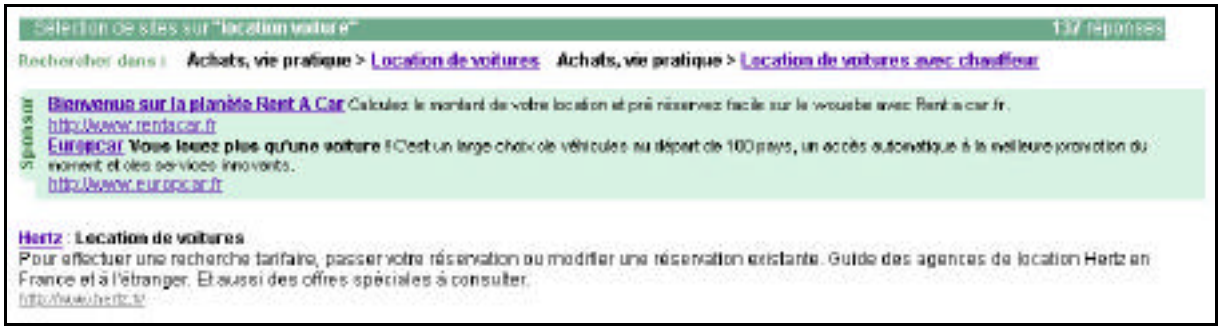
Le formulaire permet une recherche sur Voila, et un lien sur le Guide du web (l'annuaire) est proposé à droite.

La page de résultats propose les informations suivantes :

- **Services Wanadoo** :

Cette zone propose de l'information évènementielle, des contenus et services internes. Il s'agit de l'actualité (extraits de la base des dépêches d'actualité de l'AFP), des dossiers (réalisés par les documentalistes du Guide) et les espaces thématiques. Le nombre de liens proposés est limité : 1 évènement, 3 services et 1 dépêche d'actualité, soit 5 liens au maximum (sur Voila, il est affiché cependant 1 service de plus, soit 6 liens en tout s'ils sont disponibles, bien sûr).

- **Sélection de sites :**



On trouve tout d'abord, dans cette zone, une ou deux catégories (qui contiennent le(s) mot(s) demandé(s)) et proposés en dessous. Les liens vers les catégories ont, en revanche, disparu des résultats "sites web" proposés dans cette rubrique.

Deuxième zone affichée : un ou deux liens promotionnels (offre "Pole Position" de Wanadoo Portails). A noter que si deux liens promotionnels sont affichés, la zone "Sélection de sites" proposera autant de sites en moins. En clair, la zone proposant 5 liens en tout, dans laquelle on peut trouver :
- Aucun lien promotionnel et 5 liens issus du Guide ;
- 1 lien promotionnel et 4 liens issus du Guide ;
- 2 lien promotionnels et 3 liens issus du Guide.

Tout le Web francophone :

La zone "Tout le Web francophone" propose, enfin, 5 liens issus du moteur. Si un seul lien issu du Guide est proposé, ce sont 9 liens "moteurs" qui seront affichés, la finalité étant d'arriver à 10 liens pour l'ensemble "liens promotionnels + guide + moteur". Pour chaque lien, le titre de la page, un "snippet" (un extrait du contenu textuel de la page, "autour" du mot demandé) et l'url sont proposés. Il est cependant possible d'afficher plus d'information à l'aide du menu "Personnaliser" (à droite du formulaire, en haut d'écran).

L'ordre d'affichage est donc bien différent par rapport à la version précédente, notamment sur Voila, qui affichait jusqu'à maintenant TOUS les sites de l'annuaire PUIS TOUTES les pages du moteur. A cette époque, les liens "moteurs" étaient donc souvent relégués très loin dans les pages de résultats, ce qui n'est plus le cas. Une bonne nouvelle pour les spécialistes en optimisation de pages !

Revenons à la fonction "Personnaliser" : elle permet d'activer ou de désactiver le correcteur orthographique (voir plus loin), de modifier le nombre de réponses par page (10, 20, 30, 40 ou 50 liens) et enfin d'afficher plus ou moins d'informations par lien : clustering (choix activé par défaut, un seul lien par site), affichage des titres uniquement, possibilités d'obtenir un extrait textuel des liens, marqueur qui indique où se trouve le mot demandé dans la page. Ces deux choix étaient, auparavant, proposés par défaut à côté des liens. Ils sont maintenant disponibles en option. On peut penser que c'est une bonne chose, car cela aère considérablement la page de résultats.

Le choix "Afficher une seule réponse par site" ne semble cependant pas fonctionner et semble surtout désactiver le filtre adulte, activé par défaut... De plus, à chaque validation d'une nouvelle configuration, on reçoit un message : "Désolé, aucun document ne correspond aux termes de votre recherche. Solutions possibles pour trouver des réponses à votre requête : >> Votre recherche est trop restrictive, utiliser des mots-clé plus généralistes >> Vos mots clé sont mal orthographiés"... Bizarre, peut-être un bug de jeunesse.

Autres nouveautés sur la page de résultats :

* Un "filtre adulte" est activé par défaut (désactivé par défaut sur Voila, les deux populations visées étant différentes). Un bouton permet de l'activer ou de le désactiver sans problème. Il semble fonctionner de façon assez pertinente. Ce filtre permet d'enlever, aux dires de Wanadoo, également pas mal de pages de spam (pages satellites, sites identiques sous des noms de domaine différents, etc.) ou de "contenus non identifiés", qui peuvent être intéressants mais sans certitude. Sur Wanadoo, ces pages sont éliminées par défaut et sur Voila, disons que l'outil "prend le risque que ces pages soient intéressantes", puisque le filtre est désactivé par défaut. Là encore, les populations différentes sur Voila et Wanadoo sont les raisons de ce choix.

* Un correcteur orthographique est activé par défaut également. Si vous tapez "imobilier", une option "Orthographe voisine ? immobilier" sera proposée. Très utile...

* Des "Related Searches" (expressions connexes) sont affichées : le mot clé "immobilier" va suggérer "location immobilier", "crédit immobilier", "prêt immobilier". 3 expressions sont proposées au maximum. En revanche, l'option disparaît si l'on désactive le filtre adulte, sans réelle explication. Il s'agit de "Related Searches" au sens strict du terme, c'est-à-dire basés sur les expressions contenant le terme demandé et ayant été le plus souvent demandées lors des 4 semaines précédentes (délai approximatif : les mises à jour des "Related Searches" devraient être faites à l'avenir tous les mois au maximum, voire sur des délais plus courts). Pas d'analyse sémantique, statistique ou linguistique (de type Exalead) pour l'instant. Ces "Related Searches" ont apparu un temps sur Wanadoo, puis ont disparu. Elles devraient réapparaître d'ici peu, dès qu'un petit problème technique aura été réglé...

En ce qui concerne la syntaxe d'interrogation, elle est rappelée dans la partie "Aide en ligne > Trucs et astuces". La plupart des fonctionnalités déjà possibles dans la version précédente sont disponibles :

- **title:** la recherche s'effectue sur le titre du document.
- **anchor:** les termes de votre requête doivent être contenus dans le texte d'un lien hypertexte. Mais est-il normal que ce type de requête fournisse des liens émanant du Guide ?
- **keywords:** les termes de votre requête doivent être contenus dans la balise "meta keywords" de la page source. Même remarque que pour anchor:.
- **desc:** les termes de votre requête doivent être contenus dans la balise "meta description" de la page source. Même remarque que pour anchor:.
- **link:** le résultat doit contenir un lien vers l'URL spécifiée. Vous recherchez dans ce cas les pages qui contiennent un lien vers l'URL que vous inscrivez dans la boîte de recherche : link:www.abondance.com donnera toutes les pages qui contiennent un lien vers le site Abondance. D'après nos tests, une requête comme :
link:www.abondance.com -domain:www.abondance.com
qui permet d'obtenir toutes les pages contenant un lien vers le site Abondance en en excluant les pages de ce site, ne semble pas fonctionner. Dommage... Même remarque, sinon, que pour anchor: est-il normal que ce type de requête fournisse des liens émanant du Guide ?
- **alt:** recherche d'un terme dans une balise image. Intéressant pour rechercher des images. Testez la fonction sur "alt:ferrari", par exemple. Même remarque, cependant, que pour anchor:.
- **domain:** le terme de votre requête doit se trouver dans le nom de domaine.
- **url:** le terme de votre requête doit se trouver dans l'URL. Cette fonction, ainsi que la précédente, fonctionne, en revanche, sur le Guide et le moteur, ce qui est normal. En revanche, lorsqu'il y a peu de réponses, la fonction de clustering (1 seul lien par site), pourtant validée dans les préférences, semble ne plus fonctionner : on trouve alors beaucoup de liens issus du même site.

Précision : tous les petits problèmes rencontrés ci-dessus semblaient être en cours de réparation au moment de la sortie de cette lettre... Ils n'apparaîtront peut-être plus lorsque vous lirez cet article... Les petits bugs de jeunesse...

Notons enfin une fonction "Recherche sur le Guide mondial" qui prend en compte l'Open Directory, avant que Voila ne développe son propre moteur mondial (en projet). En revanche, effectuer des recherches sur l'Open Directory par l'intermédiaire de Voila ou de Wanadoo permet de contourner le probleme actuel des urls avec caractères accentuées sur le site http://www.dmoz.org/ (lorsqu'on tape "astérix" et qu'on clique sur l'une des catégorie proposées, on arrive irrémédiablement sur un erreur 404)...

Enfin, pour terminer, notons que les pages de résultats de Wanadoo et de Voila sont aujourd'hui, à part quelques petites différences dans la section "Services", quasi identiques. Les remarques faites ci-dessus pour l'un sont donc valables pour l'autre.

Le Guide Voila : un nombre de sites stable

Le Guide, accessible depuis le lien de la page d'accueil (http://recherche.wanadoo.fr/), propose un annuaire de sites pour enfants, qui existait depuis très longtemps (il était déjà proposé à l'époque

de QuiQuoiOù), mais qui est mieux mis en valeur et remis à jour.

La section "Nouveaux sites" propose les sites nouvellement intégrés dans l'annuaire. L'outil propose environ 63 000 sites dans 14 000 catégories. Le nombre de sites ajoutés dans le Guide compense le nombre des sites supprimés, ce qui explique le nombre de sources d'information annoncé qui reste stable depuis de nombreux mois. A noter que l'indexation des sites dans l'annuaire est aujourd'hui quotidienne. Un site évalué et accepté par les documentalistes est présent le lendemain, sauf incident, dans la base de données et donc visible sur le Web en 24 heures.

Un projet d'offre de "Modification Express" des sites déjà dans l'annuaire et n'étant pas passé par l'offre de Soumission Express (soumission payante) est également à l'étude et devrait voir le jour d'ici à la fin de l'année : il permettrait, pour une somme moins importante que celle de la Soumission Express, de demander des modifications de la fiche descriptive du site déjà présent dans l'annuaire. Plus d'infos certainement sur ce projet à la rentrée prochaine...

La soumission payante, quant à elle, devrait rapidement devenir obligatoire pour les sociétés, donc dans la branche "Entreprise, Economie", notamment (comme sur Yahoo! France depuis quelques jours). Certainement à la rentrée. A noter que certaines nouvelles contraintes ont été rajoutées dans les conditions générales de vente, pour éviter certaines tentatives de spam (eh oui, même en payant, il y en a qui spamment !!). Peuvent donc soumettre uniquement les sites qui répondent à ces conditions :

- Le site doit comporter une version française ;
- Son contenu doit être conséquent, mis à jour régulièrement si nécessaire, et ne pas enfreindre ou ne pas violer les droits d'un tiers, ni proposer de liens vers des contenus susceptibles de le faire ;
- Le site ne doit pas proposer de contenus illégaux ou jugés inappropriés par le documentaliste en charge du dossier. Seront notamment refusés les contenus pornographiques ou réservés aux adultes, diffamatoires, racistes, violents, ou proposant des liens ou des images vers de tels contenus ;
- Le site doit contenir un nombre de pages suffisant et un contenu permettant aisément de déterminer une catégorie ou sous-catégorie d'appartenance ;
- Les sous-parties d'un site ne sont pas référencées à moins qu'elles proposent un contenu suffisamment conséquent permettant de déterminer une catégorie d'appartenance différente de celle du site principal ;
- Le site doit fonctionner 7 jours sur 7 et 24 heures sur 24 ;
- Le site, au moins pour sa version française, ne doit pas être en construction ; tous les liens doivent fonctionner et les pages doivent se charger rapidement ;
- Le site ne doit pas être un site miroir ou rediriger vers un autre site.
- Le site doit être compatible avec les principaux navigateurs utilisés au moment de la soumission.
- Si le site nécessite une inscription et un mot de passe, le soumissionnaire doit être en mesure de fournir sur demande des éléments d'identification permettant de tester le site et ce à tout moment suivant le paiement de la prestation.
- Le site ne doit pas faire la promotion de sectes, de mouvements spirituels ou de thèses polémiques et choquantes.
- Le site ne doit pas proposer d'activité dans le domaine des casinos virtuels.
- Le site ne doit pas uniquement proposer de contenus tels que : un curriculum vitae, un book, une annonce de vente émanant d'un particulier.
- Le site ne doit pas être une url doublon : inscription multiple d'un même site avec des extensions différentes ou inscription de plusieurs pages d'un même site déjà référencé.

On a effectivement dans ce texte une liste quasi-exhaustive des type de tentatives de spam sur un annuaire...

Moteur de recherche (KE) : peu de changements

Le moteur de recherche de Voila / Wanadoo (nom de code : KE) est assez stable depuis un an. En gros, il n'a pas été l'objet de beaucoup de modifications depuis ce laps de temps. Les articles écrits sur ce sujet depuis 12 mois dans cette lettre restent donc globalement valables :

Voila moteur : les nouveautés et les projets :
http://abonnes.abondance.com/archives/acturech/0112.html

Voila : un point sur l'offre "Pôle Position" :
http://abonnes.abondance.com/archives/acturech/0203.html

Voila.fr et Nomade.fr : un point sur le référencement payant :
http://abonnes.abondance.com/archives/acturech/0201.html

Actuellement, le moteur KE "encaisse" aux alentours de 3,2 millions de requêtes les jours de semaine (3 millions en moyenne si on prend en compte les week-ends), mais peut connaître des pointes à 3,5, voire 4 millions de requêtes/jour. L'index du moteur propose 60 millions de pages francophones, le "refresh" (mise à jour de l'index) est effectué toutes les semaines pour certains sites, et tous les mois pour l'index général, mais ce délai peut fluctuer. Le délai entre la soumission d'une nouvelle URL et sa prise en compte par le moteur est d'environ un mois actuellement, mais sans garantie, ces délais sont très variables à l'heure actuelle.

Enfin, parmi les projets en cours : une internationalisation du moteur sur l'Europe notamment et des "Wanadoo Ville", portails spécialisés pour certaines villes françaises.

Conclusion

Le nouveau "look" des pages de résultats de Voila et de Wanadoo est beaucoup plus clair, plus lisible, à notre sens, qu'auparavant. Il remet en valeur également le moteur de recherche en plaçant de façon obligatoire des résultats émanant de son index en première page ce qui n'était pas le cas auparavant.

Il ne reste donc plus qu'à corriger quelques petits bugs dans les fonctionnalités avancées et les préférences d'interface et ce sera parfait. Il sera alors temps de s'attaquer au chantier de la pertinence des résultats "moteurs", des outils comme Google, Wisenut, Teoma et Exalead ayant beaucoup progressé dans ce domaine alors que Voila donnait l'impression de stagner un peu depuis un an. Mais, au vu de cette nouvelle refonte du "look" des pages, il semblerait que l'outil de recherche soit sur la bonne voie...

Webchercheurs : un partenariat Webhelp / Voila / Wanadoo

Retour au sommaire de la lettre

Connaissez-vous le service "Web Chercheurs" (http://webchercheurs.servicesalacarte.voila.fr/ et http://webchercheurs.servicesalacarte.wanadoo.fr/) ? Pas sûr, car le lien qui fournit l'accès à ce nouveau service n'est pas facile à trouver sur les pages de résultats de Voila et de Wanadoo (regardez bien, en bas de page, le lien intitulé "Des experts cherchent pour vous"). Pourtant, le système est plutôt efficace et ressemble fort à ce que propose WebHelp (normal, c'est à eux que Voila sous-traite en fait le service) : vous saisissez votre adresse e-mail et votre question, vous indiquez si vous acceptez les sites en anglais et les sites persos, puis vous choisissez votre système de paiement :

* Par le système W-HA qui débite automatiquement l'abonné Wanadoo (mais le système W-HA fonctionne également avec les abonnés Tiscali et Club-Internet) de 1,89 euros sur la facture de son fournisseur d'accès.

* Par le système AVA, où le débit s'effectue par l'appel d'un numéro audiotel qui fournit un code à saisir dans l'interface Web (1,349 euros l'appel puis 0,34 euros par minute)

Bref, dans les deux cas, poser une question en ligne coûte un peu moins de deux euros, soit 12 francs environ, ce qui est assez raisonnable (non ? Si...).

Pour plus d'information sur ces deux modes de paiement, connectez-vous ici : http://www.leskiosques.com/

On peut noter que, pour l'instant, 95% des achats s'effectuent par le mode W-HA et 5% avec AVA. L'offre est disponible sur Wanadoo et Voila, mais 90% des personnes qui viennent poser des questions viennent de Wanadoo et 10% de Voila. Le profil des Wanadiens, certainement moins "pointus" dans leurs recherches que les Voilanautes, y est certainement pour quelque chose...

250 questions par jour

Le service, techniquement et humainement géré par WebHelp, répond à l'heure actuelle à 250 questions par jour environ, ce qui montre que les internautes sont prêts à payer une somme, modique certes, pour obtenir une réponse à leur question. Au vu du peu de visibilité des liens vers ce service, on peut estimer que ce nombre de questions traitées est plutôt bon. 25 "Webwizards" (noms des opérateurs chargés d'effectuer les recherches sur le Web) travaillent sur ce service et sont basés à l'étranger (4 Webwizards y sont également chargés de travailler sur le site www.webhelp.fr, devenu payant depuis quelques mois). Les meilleurs Webwizards de l'équipe sont dédiés au service "Webchercheurs".

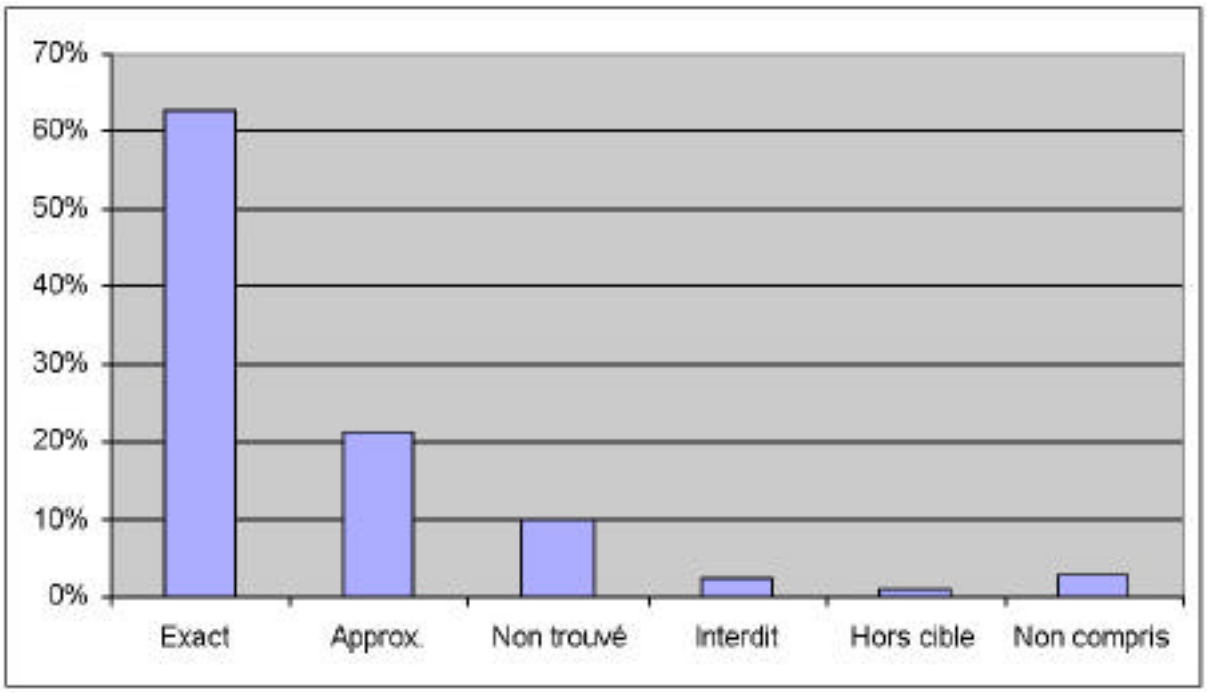
Les informations parues dans la lettre d'avril 2001 sur WebHelp restent encore valables (http://abonnes.abondance.com/archives/acturech/0104.html), les méthodes de travail restent à peu près les mêmes, sauf pour l'outil informatique utilisé qui est beaucoup plus performant aujourd'hui.

En règle générale, les Webwizards traitent 4 questions par heure. Le but est de fournir 2 à 3 liens par réponse. Aujourd'hui, 10% des réponses proposent 1 lien, 20% proposent 2 liens et 70% proposent 3 liens et plus.

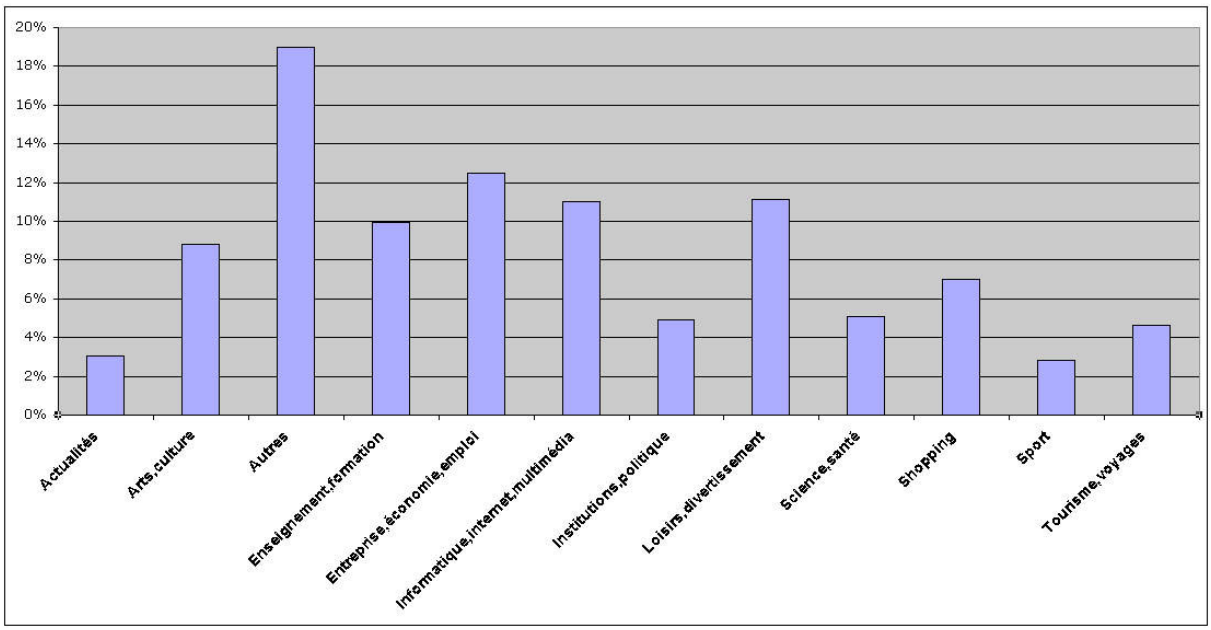
WebHelp a été surpris par la complexité des questions qui sont posées depuis le lancement. Il faut dire que le fait de payer élimine certainement de nombreuses questions un peu "bidons", ce qui n'empêche pas quelques "perles" (voir bétisier plus loin).

Un outil d'estimation de la qualité des réponses a été mis en place. Ainsi, sur les premières semaines d'exploitation, une réponse exacte a été fournie à 53% des questions (niveau 1) et 15% (niveau 2, Webwizards plus "costauds"), soit 68% des questions, ce qui est plutôt bien. 19% des réponses ont été considérées comme "approximatives", 5,6% des questions n'ont pas trouvé de réponse (pas trouvé sur le Web), 3,5% des questions étaient "interdites" (sexe, pédophilie, piratage, etc.) et 3,5% ont été considérées comme incompréhensibles. Si, après 24 heures, le

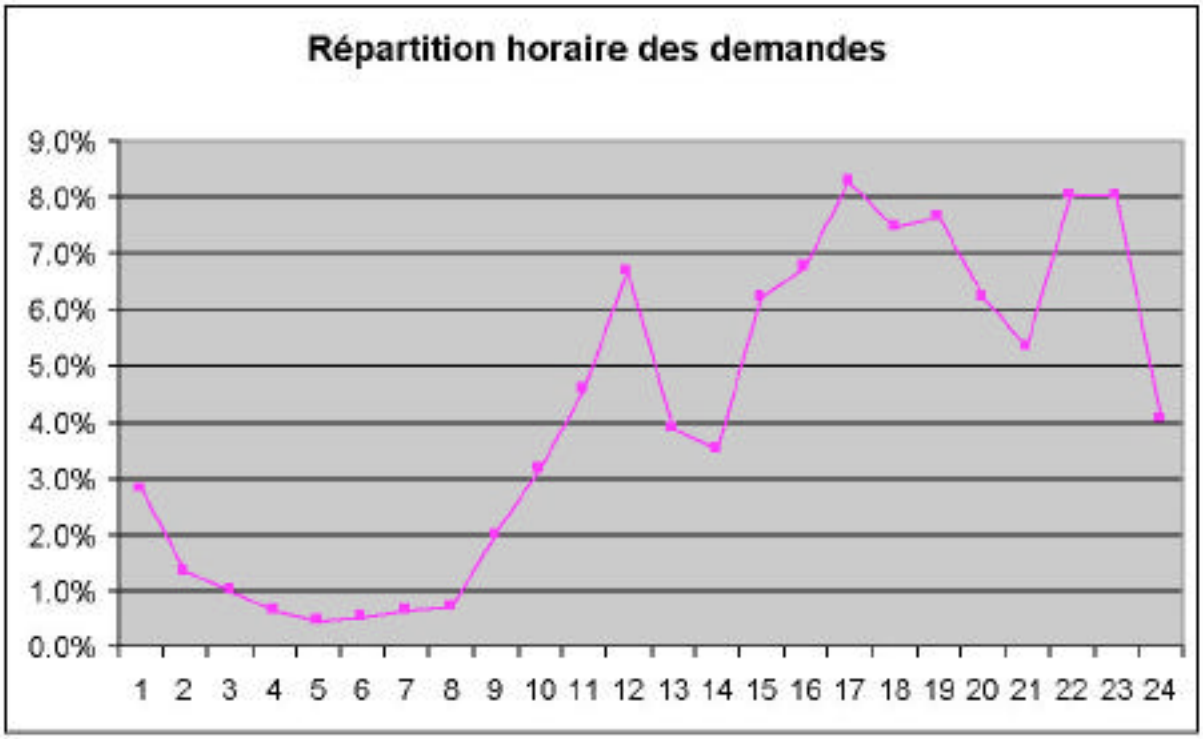
Webwizard n'a pas trouvé de réponse, il propose une autre recherche gratuite à son client. Ainsi, 15% des recherches sont effectuées à titre gratuit et 85% sont payantes.



Les catégories thématiques des recherches sont les suivantes :



Les pics de consultation se situent entre 16 h et 23 h :



Le service payant WebHelp (http://www.webhep.fr/), qui occupe, comme nous l'avons dit précedemment, 4 WebWizards à temps plein, répond à une cinquantaine de questions par jour.
La répartition des paiements est celle-ci : Audiotel : 24% et abonnement : 76%.
Ce sont plutôt des hommes (67%) qui utilisent ce service.
La typologie des demandes est la suivante :
- Informatique, Internet : 23%
- Entreprise, économie : 18%
- Histoire : 6%
- Tourisme, voyages : 5%
- Vie pratique : 5%
- Enseignement : 5%
- Sujets de société : 5%
- Actualtés, médias : 4%
- Administrations, politique : 4%
- Loisirs, sorties : 4%
- Santé, médecine : 4%
- Arts, culture : 3%
- Science, recherche : 3%
- Shopping : 3%
- Emploi : 2%
- Sport, plein air : 1%
- Villes, régions, pays : 1%
- Autres : 3%

Voici, pour information, une liste des questions représentatives sur le service "Web Chercheurs". Vous vous apercevrez ainsi de la diversité des questions posées :

** Je recherche des exemples de texte pour faire part de mariage. (exemples de textes originaux.) pouvez-vous m'aider, merci*

** *********, candidat aux legislatives 2002 dans le maine et loire. on a envoyé *******-legislatives.com mais je ne trouve rien. merci de m'aider. sylvette*

** souraya serait le nom d'une deesse? quelle est l'origine de ce nom ?*

** L'AGEFIPH et l'Etat : quelles sont les dernières évolutions en matières d'obligation d'emploi et d'aide des travailleurs handicapés dans la fonction publique ?*

** Je recherche une chambre chez l'habitant à Miami en Floride*

** JE RECHERCHE LE DRIVER COMPATIBLE XP POUR UNE CARTE TV PCI DE MARQUE TYPHOON (BT878)*

** M. ***** actuel Recteur de l'Ac. de Rennes aurait été mis en examen suite à ses "activités" précédentes (Recteur en Corse) . Merci à l'avance pour une information-réponse.*

** Sur quels sites puis-je trouver des chiffres concernant le marché des plastiques et plastiques extrudés?*

** Mon fils achete une moto de cross d occasion et la revend huit mois plus tard en specifiant vendu dans l etat. 3 semaines apres l acheteur se manifeste en evocant le vice cache du bouchon de vidange soude par le precedent proprietaire .et me somme de prendre les reparations a ma charge 500 euros. ou de m assigner devant le tribunal . merci d avance*

** Recherhe ***** ******, Age 67ans Service militaire: 12 RCA Meknes - MAROC 1956 a 1958*

** Je recherche des informations dans la presse sur la désinformation, les informations truquées ou détournées, la création et les conséquences des rumeurs sur la société*

** Je voudrais avoir le nom des sites proposant l'évolution du vélo de son début à aujourd'hui, s'il vous plait.*

** Recherche site sur la marque dap.(pieces de karting). FREIN ET CHASSIS DAP.*

** Je recherche les sites où je pourrais acheter des figurines de dragons, comme celles qui sont sur le site "fugur-in.com". ais je voudrais d'autres modèles. Merci*

** Je recherche des sites montrant des examens gynécologiques*

** Fabrication des cocottes en papier*

** QUEL EST LE FONCTIONNEMENT DE LA TVA INTRACOMMUNAUTAIRE*

** Clinique ou centre d aide a la procreation artificielle a geneve en suisse;*

** Quel facteur alimentaire a fait que la dentition des populations pré-moyenageuse etait en bien meilleur etat que celle des populations du moyen-age?*

** Je cherche un revendeur d'echelle pour une échelle en bois de 2.50 metres eventuellement echelle de meunier sur le 06 de preference.merci*

** Les conséquences de la maladie de Parkinson sur la conduite automobile*

** Je cherche le délégué aux rapatriés d'Algérie du nouveau gouvernement*

** Je voudrais une lettre type pour envoyer mes condoleance à une amie qui vient de perdre son mari agé de 32ans merci*

** Je cherche l'image de la pub omo micro*

** Je cherche differentes pages sur le pedigree des pit bull c.à.d qui était la mére le grand pére etc.merci*

** Recherche clinique spécialisée privée pour personne psycho maniaco dépressive*

** Je cherche des sites qui parlens d'alain dewerpe*

** A quoi correspondent les noms suivants : peplin b.s duchownego ou duchovnego*

** Comment se sentir bien dans sa tete et dans sa peau apres un arret de tabac de 1 mois et surtout ne pas craquer sur l alimentation*

** Comment savoir si les adresses e mail de mon cousin au usa est correct car ecriture illisible*

** J'aimerais connaitres les mysteres du triangles de bermudes. les profondeurs de l espaces. si il y a une vie ailleurs*

** Souhaite logiciel d'anesthesie par inhalation "gasman"*

** Tout ce qui ce rapporte à un lac de Richwiller appelé "club de la plage" ou dans le passé "Mulhouse plage" (photos articles ...)*

** J'aimerai trouver le site page perso le coscro, est il possible de le trouver par voila?*

Et, pour finir sur une note amusante, voici le bétisier du service depuis son lancement (après de longues hésitations, j'ai enlevé les questions un peu trop... osées ;-)) :

** Je veux savoir si l'acteur brade gorton et homo et savoir ou il habitte et savoir si il a quelqu'un en se moment merci*

** Puis-je avoir accès à l'imposition de mon ex-femme Catherine **** née le 13/03/1957, vivant à LOUVECIENNES (78) avec M Jean *********

** Mon mari est parti depuis 1991 nous ne sommes pas divorcés devrais-je participer aux frais d'obséque à son décès et réciproquement y a t il obligation ?*

** Je veux faire une surprise à mon ami, lui faire ma déclaration dans les air cad écrire je t'aime dans le ciel par l'armée de l'air.*

** Je cherche des solutions pour avoir une poitrine poilue.*

** G une mauvaise formation de la verge et g des testicules trop petites*

** Comment enlever du sicaflex sur du komusssal?*

** Quelles sont les dimensions du col de l''utérus chez une nullipare ?*

** Je voudrais savoir comment faire pour effacer mon casier judiciaire car je travaille en CDD dans une mairie et si mon casier n'est pas vierge il ne m'embauche pas en CDI*

** Je voudrais un site ou on peut assosier 2 races de chien pour voir ce que ca donne*

** Recherche un navire petrolier a vendre*

** Je recherche le titre d'un film avec jacques villeret ou il a un role de gros degueulasse il porte un slip blanc tache devant et derriere*

** Trouver de l'argent à n'importe quel prix*

** Dans quelle rubrique peut-on proposer ses services (gratuitement) pour garder une maison avec piscine région de Toulon pendant les vacances scolaires - ref. merci*

** Je connais quelqu'un qui fraude dans sa declaration de revenu et voudrais le denoncer . a quel organisme m'adresser ?*

** RENCONTRANT UN PROBLEME DE VOISINAGE INSULTE VERBAL ET GESTE OBSENE NOUS AVONS DEJ0 PORTE PLAINE MAIS RIEN N'a abouti de plus le voisin connait personnellement l'adjudant de gendarmerie. Je possede une protection juridique vie privee que faire???*

** QUEL EST LE MEILLEUR RAPPORT QUALITE-PRIX POUR L'ACHAT D'UNE PISCINE OCTOGONALE ?*

** Je recherche des endroits où acheter des insectes commestibles pour les hommes (boutiques ou restaurants).*

** J'aimerai trouver la partition au piano de: "Sur un prélude de Bach" chanté par Mauranne,mais je ne sais pas qui est le musicien.*

** Je cherche des informations les plus diverses possibles sur les sites traitant des différents modes de crochetage de serrure (si possible avec explications).*

** J'ai 33 ans durant le mois de aout des cousins sont partis au brésil et on été assassinnée il était 6 et je voudrait bien trouver un site avec leur photos de cadavres*

** Recherche compagnie assurance automobile pour contract résilié pour alcoolemie*

** Je cherche des produits amincissants pour hommes, j'ai de la graisse localisée au niveau des péctauraux & je cherche 1 produit visant à réduire cette masse graisseuse.*

** J'ai une affaire qui se juge jeudi prochain et je cherche un autre avocat car la mien est completement saoul et ne fait que me demander des honoraires sans rien m'expliquer de l'affaire. j'habite issy les moulineaux et je cherche un avocat specialise en droit des personnes*

** J'aimerai savoir fabriquer des saucisses de canard gras sans porc*

** Peut-on avoir une pièce en apesenteur chez soi?*

** LE DICTIONNAIRE DES PRENOMS KURDES*

** Comment annuler tout ce que j ai pu entrer comme informations sur internet*

** Je cherche un medecin lyonnais qui utilise un laser co2 pour traiter un pb de folliculite pilleux sur le nez merci d'avance pour votre reponse*

Ah, au fait, j'avais oublié... J'ai testé ce nouveau service sur deux questions. A vous de juger de la pertinence des réponses :

QUESTION 1 : "Combien existe-t-il d''encyclopédies en français sur le web ?"

REPONSE 1 : "Nos experts ont effectué une recherche sur internet et ont trouvé les réponses suivantes :
Lyc-llg.scola.ac-paris : http://lyc-llg.scola.ac-paris.fr/liens/dictionnaires/#scientifique
Web.ccr.jussieu : http://web.ccr.jussieu.fr/urfist/cerise/lexdico.htm
Webencyclo :
http://www.webencyclo.com/Inscription/prehome_Flechage1.asp?IDFORM=%A3%A3&FromHome=irncRfr4&idDoc=&idm="

QUESTION 2 : "Quelle est la taille du web ?"

REPONSE 2 : "Nos experts ont effectué une recherche sur internet et ont trouvé les réponses suivantes :
Abondance : http://www.abondance.com/docs/question10.html
Adbs : http://www.adbs.fr/adbs/sitespro/lardy/chap4.htm
Remarque: Pour trouver plus rapidement l'information demandee sur la page envoyee, veuillez appuyer simultanement les touches Ctrl et F et entrer le mot « taille » dans la boite de recherche
C.asselin : http://c.asselin.free.fr/french/webenchiffre.htm
Remarque: Merci de cliquer sur la rubrique "Taille du web-web invisible"."

DeepIndex : un nouvel outil de recherche

Retour au sommaire de la lettre

La société VirtuelPub a dernièrement mis en ligne le moteur de recherche Deepindex (http://www.deepindex.com/). Pourquoi mettre en ligne un nouvel outil à une époque où la tendance est plutôt au resserrement de l'offre, ce marché se portant moyennement bien, subissant de plein fouet l'apathie du marché publicitaire ? Une bonne occasion pour essayer d'en savoir plus sur ce nouvel acteur de la recherche d'information en interviewant Gilbert Wayenborgh, responsable de la société VirtuelPub.

- Gibert Wayenborgh, pouvez-vous nous présenter en détail le nouveau moteur DeepIndex ?

Le moteur de recherche DeepIndex se présente sous deux formes. La forme la plus classique avec un formulaire de recherche, et une forme plus assistée, avec des thèmes prédéfinis, et des mots clés prédéfinis dans le thème que nous appelons Miniportail. L'utilisateur peut rechercher un mot clé ou des combinaisons de mots clés et utiliser les opérateurs booléens. Enfin pour affiner ou limiter sa recherche, il peut rechercher dans le titre ou dans le corps du texte.
Pour terminer, nous reprenons également le contenu de la base de données de France-Sites (http://www.france-sites.com/), contenant pas moins de 50.000 sites francophones. Celle-ci sera mise en ligne d'ici quelques semaines.
L'utilisateur dispose à tout moment d'une interface permettant de discuter avec les membres de DeepIndex pour critiquer par exemple nos résultats de recherche, ou encore pour nous demander une recherche spécifique. Une équipe est prévue à terme, mais tout au plus 3 à 4 personnes. Il est inutile de charger financièrement l'exploitation et de faire les mêmes erreurs que d'autres outils disparus ont commis.
L'indexation est réalisée par un crawler appelé DeepIndexer qui scrute quelques annuaires et qui suit ensuite tous les liens présents. La technique qui est actuellement derrière DeepIndex est un applicatif "Open source" (Aspseek http://www.aspseek.org) auquel nous ajoutons nos outils internes et systèmes ou plutôt méthodes d'exploitation. L'indexation d'un site se déroule en général en plusieurs phases. La première phase indexe une dizaine de pages, lors d'un deuxième passage les liens internes du site sont indexés. Le delais moyens pour une prise en compte payante est moins d'une semaine. En version gratuite il faut attendre le "refresh", qui est actuellement trimestriel. Mais la version que nous développons avec Multi Vision International doit nous permettre une indexation beaucoup plus rapprochée.
Le robot suit les directives robots.txt et est calibré pour ne pas saturer les sites fragiles.
Techniquement il existe actuellement 2 serveurs principaux, dont l'un est hébergé en Aquitaine, et un deuxième à Montréal.

- Quelques chiffres (trafic actuel, configuration technique, taille de l'index, etc.) ?

Le trafic actuel est, vous vous en doutez, encore peu significatif, l'outil ayant été lancé en juin dernier. Néanmoins la progression est rapide et les partenariats établis nous permettent de prévoir des chiffres significatifs pour septembre. La configuration technique est basée sur deux serveurs avec un applicatif "Open source". La taille de l'index est de 2.000.000 d'urls, ce qui peut paraître modeste par rapport aux chiffres annoncé dans vos colonnes. D'ici la fin de l'année nous estimons que la taille devrait tourner autour de 20.000.000 d'urls, essentiellement francophones, mais nous préparons également d'autres serveurs spécifiques tel qu'un serveur hispanique..

- Quel a été l'investissement financier et humain pour mettre en place un tel moteur ?

L'investissement a été essentiellement humain avant d'être financier. La configuration de base a été montée en Mars dernier, afin d'expliquer le projet et son intérêt à des partenaires. Les réels

investissements ne font que commencer par une version DeepIndex redéveloppée et architecturée différemment, afin de pouvoir évoluer plus rapidement pour nous hisser dans les majors francophones rapidement. A l'heure actuelle 30 personnes sont impliquées dans le projet. Ce sont essentiellement des profils ingénieurs systèmes Unix/NT, dba, commerciaux, et 5 webmasters.

- Qui sont les partenaires et les outils de développement avez lesquels vous avez travaillé ?

Il y a quatres partenaires principaux :
* Cristal-Trace, l'hébergeur qui nous a suivis dès la première heure.
* Networldmédia, le sponsor principal et qui assure la promotion avec beaucoup de talent au Canada.
* Multi-Vision International et plus particulièrement son département "recherche et développement", qui travaille sur la prochaine version applicative. Cette équipe a remporté la médaille d'or de l'OMPI (l'Organisation Mondiale de la Propriété Intellectuelle) en novembre 2001 pour une invention concernant la recherche d'informations en temps réel "bourse et sites de news" et la technologie SMS.
* virtuelpub.com qui se charge du référencement et de la promotion du moteur en France.

- Pouvez-vous nous parler de la fonction "rédiger une critique" ?

Cette idée m'est venue en consultant les nombreux forums ou listes de diffusion dédiés au monde des outils de recherche tant francophones qu'américains. Cette fonction permet à l'utilisateur de nous indiquer un site qui ne satisfait pas sa demande et qui ne correspond pas à sa demande. Accessoirement cette fonction permet également de dénoncer le spamdexing. Mais attention, les "critiques" sont systématiquement traitées manuellement et vérifiées.
Nous recevons la critique par e-mail. Nous exécutons d'abord la requête, puis vérifions le contenu du site en question dans l'ensemble.
Si la critique est mal formulée, ou mal intentionnée, nous ne donnons aucune suite. Si au contraire elle est justifiée, nous envoyons d'abord un mail au Webmaster concerné lui laissant une semaine pour nous répondre, puis le cas échéant le site est déclassé.

- Lorsqu'une recherche n'a rien donné, le message "Pour enrichir notre index, nous pouvons rechercher pour vous" s'affiche. Pouvez-vous nous en dire plus ?

Les recherches sont souvent complexes à formuler pour les internautes et notre index est récent. Certaines requêtes restent de fait encore sans réponse. L'utilisateur peut alors nous indiquer ce qu'il cherche, et nous cherchons alors pour lui. Nous indexons alors quelques sites ou pages contenant ce qu'il cherche. La réponse est envoyée par e-mail.
Pour anecdote, certains utilisateurs cherchent parfois des proches, un peu comme une certaine émission de télévision. DeepIndex n'est pas perdu de vue ;-).

Le site semble assez lent. Je me trompe ?

Oui. Les recherches sont pratiquement aussi rapides que les majors. Ce qui est possible par contre à certains moments aux alentours de 6 heures du matin par exemple, nous compilons l'indexation courante pour l'intégrer dans la base de données générales. Ce "process" est lourd et le serveur peut alors présenter des lenteurs. Ce problème devrait se régler assez rapidement lorsque nous aurons terminé certains développement.

Quels critères de pertinence utilisez-vous ?

Lors de l'indexation nous utilisons l'ensemble du texte d'un page (Titre, Meta, Body et la position d'un mot dans cette page et le nombre d'occurrences.), lorsque l'utilisateur effectue sa recherche il peut affiner celle ci en indiquant quel critère il souhaite rendre important. Pas d'indice de popularité pour l'instant. Le critère de popularité est controversé et n'apporte pas forcement une information pertinente. C'est aussi un critère trop facilement manipulable, mais vendable peut-être ?

- Le moteur a-t-il des critères bloquants (pages dynamiques ? Flash ? Javascript ? Frames ? Autres ?)

Toutes les pages sont indexées, cependant nous avons volontairement limité le poids des pages. Bien entendu le Flash, le jJvascript et les cgi sont exclus pour le moment pour des raisons techniques. Cependant le PHP et l'ASP sont quant à eux indexés normalement jusqu'à une certaine limite.

- Allez-vous utiliser une technologie de liens promotionnels (Overture, Espotting ou autres) ?

Oui la technologie est répandue et efficace, l'annonceur est affiché sur un mot clé et répond donc à la demande de l'utilisateur. Notre principal partenaire Networldmedia nous fournit des liens promotionnels dans la partie moteur. Et soyons réalistes : Quel moteur ne contient pas de publicité ? Les liens promotionnels sont l'équivalent des bandeaux. Au départ nous avons contacté les trois acteurs Overture, Espotting et Networldmedia, seul Networldmedia nous a suivi. Nous les remercions par ailleurs vivement car ceux-ci nous représentent au Canada et en Espagne et proposent notre technologie à des portails d'envergures. Un premier grand portail est en cours de signature, et deux autres devraient suivre avant la fin de l'année.
Les acteurs de la publicité ont besoin d'outils de recherche au même titre que les outils de recherche ont besoin de ressources financières. Ces liens sont donc naturels et indispensables.

- Que signifie l'information "Pertinence : [0.99997]" dans les résultats ? A quoi sert-elle ?

Le degré de pertinence est calculé selon la présence du mot de recherche dans le titre, les meta tags, le body. Il est proposé uniquement à titre indicatif.

- Quelles sont les différentes solutions de référencement que vous proposez ? Proposez-vous un référencement gratuit ?

Non le référencement gratuit a fait son temps. Nous disposons de trois solutions de référencement payant en fonction du budget du Webmaster. Une version avec un numéro de téléphone surtaxé, adapté aux particuliers, une version à 50 Euro et une deuxième à 100 Euro. Sur les deux versions "Pro" nous réalisons une réplication automatique vers nos autres bases.

- Pas mal d'images ne s'affichent pas, certains liens sont cassés. Est-ce encore une version beta ?

Certains éléments sont actuellement en cours de développement pour s'adapter parfaitement à un nouveau partenaire.

- Pardonnez cette question un peu provocatrice, mais quel est l'intérêt de lancer un moteur de recherche aujourd'hui, à une époque où Google et tant d'autres ont investi un marché qui, pour sa part, semble se retrécir ? N'est-ce pas vain ?

C'est justement ce rétrécissement qui nécessite l'émergence de nouveaux outils. Mais il est vrai que la période semble peu propice et pourtant notre concept semble intéresser plus d'un. La preuve en est nos partenaires qui nous ont suivi et nous suivent encore. La croissance que nous connaissons au niveau trafic démontre également un intérêt du public. Techniquement nous n'avons rien à envier à nos concurrents et même la taille de notre index ne nous pose pas un problème majeur, hormis que nous ne pouvons communiquer sur une taille d'index de 2 ou 3 milliards de pages dont un 1/10ème est peut être consulté.

- Pensez-vous revendre la technologie DeepIndex comme Google le fait avec Yahoo! par exemple ?

Oui, d'ailleurs les premières signatures doivent avoir lieu pendant cet été mais je ne peux pas encore en parler maintenant.

- Plus globalement, quel est le modèle économique de DeepIndex ?

Le modèle économique de DeepIndex est basé en partie sur les entrées publicitaires, mais la majeure partie provient des services d'indexation rapides. Enfin moins visible pour le grand public est la vente de notre technologie et ses services associées.

- Quelles sont les ambitions de ce nouveau moteur ?

Notre ambition à très court terme est de rentrer dans le TOP 10 des outils francophones. Ensuite nous verrons comment nous positionner sur les marchés internationaux.

- Quel sont les projets à court et moyen terme pour DeepIndex ?

Nos projets à court terme se limitent à sortir rapidement notre nouveau crawler plus performant et rapide. Ceci nous permettra dans la foulée de proposer de nouveaux services à l'utilisateur et aux Webmasters tel que la ré-indexation toutes les deux heures pour des sites d'actualités par exemple. Nous prévoyons aussi quelques développements très techniques, mais il m'est impossible d'en parler ici. Il est clair cependant que DeepIndex fera parler de lui le moment venu. ;-)

Bruits et chuchotements

Retour au sommaire de la lettre

Une rubrique qui regroupe tous les bruits et rumeurs dans le (petit) monde des outils de recherche mondiaux et francophones. Rien n'est obligatoirement vérifié, mais toutes les infos sont données... de source sûre ;-)

-> L'annuaire de Voila (le "Guide") pourrait passer entièrement en soumission payante (plus de soumission gratuite possible) d'ici quelques mois. Aucune décision ne serait encore prise à ce sujet-là en interne chez France Telecom, mais il semblerait que l'idée fasse son chemin assez rapidement.

-> Le moteur de recherche HotBot devrait entièrement être "refondu" dans les mois, voire les semaines qui viennent. Qui a dit qu'il était temps ?

-> Juste après son lancement, Google aurait proposé de racheter la technologie Altavista pour 1 million de dollars. Altavista aurait refusé. A un moment donné, Yahoo! aurait également tenté de racheter cette technologie. En vain...

-> Quelques infos complémentaires sur l'épisode "Yahoo! / Google / Fast". Actuellement, Yahoo! pencherait pour la solution Fast, mais certains petits détails ne lui conviendraient pas sur cette technologie, notamment au niveau du calcul de l'indice de popularité (entre autres). Bref, des petits détails techniques qui font que l'équipe de Yahoo! a différé sa décision. Sergey Brin, le co-fondateur de Google, nous avait d'ailleurs confirmé, il y a un mois de cela, que cette décision ne serait pas prise avant la rentrée prochaine. Le but de ce délai serait de permettre à Fast de developper de nouvelles fonctionnalités correspondant parfaitement au cahier des charges demandé par Yahoo! A ce moment-là, de deux choses l'une : soit Fast démontre qu'il est capable d'apporter toutes les possibilités supplémentaires demandées par le portail, et il serait choisi. Soit il n'y arrive pas et c'est alors Google qui serait pris. Mais tout cela est, bien entendu, énoncé avec le conditionnel de circonstance, car nous ne sommes au courant de rien... Ou presque ;-)...

-> Tiscali (qui utilise Fast et Nomade.fr notamment commes outil de recherche) pourrait intéresser France Télécom et son outil Voila. Un éventuel rachat ne serait bien entendu pas sans conséquences sur les annuaires et moteurs utilisés sur les deux outils... Plus d'infos : http://www.europemedia.net/shownews.asp?ArticleID=11266
D'autre part, une rumeur insistante, a priori différente de celle évoquée ci-dessus, parlerait du rachat d'un "grand de la recherche d'information" par un autre géant du domaine. Pas de noms bien précis pour l'instant. On cherche... ;-)

En bref...

Retour au sommaire de la lettre

-> Le pourtant excellent site Web Site Garage (http://websitegarage.netscape.com/) va fermer ses portes le 15 aout 2002 au moins pour ses fonctions Register-It, GIF Lube et Tune Up. La fonction Hitometer ne fonctionnera plus que pour les abonnés AOL.

-> Altavista est devenu insensible à la casse des lettres. Des recherches sur abondance, Abondance et ABONDANCE donnent le même résultat. Si vous voulez que le moteur prenne en compte les lettres majuscules, utilisez les guillemets : "Abondance" ne trouve plus le terme abondance... "ABONDANCE" ne trouve que cette graphie, etc.

-> Sur la nouvelle version de Kartoo (http://www.kartoo.com/), qui affiche les liens Espotting, ce dernier n'entre pas en ligne de compte dans l'algorithme de classement des sites. En revanche, si un site de la carte est trouvé aussi par Espotting, alors Kartoo utilise un lien d'affiliation. Sur ce site, il est également proposé un petit symbole triangulaire sur les boules qui donne accès à des fonctions de type "recherche sur ce site" "sites semblables" "ajouter ce site aux favoris", etc.

-> Altavista propose une page intéressante pour les webmasters à l'adresse : http://www.altavista.com/webmaster

-> Quelques infos sur l'Open Directory (partie francophone) :
Nombre de sites classés en World/Français
30 juin 2001: 43 523
27 juin 2002: 82 621
Le français est actuellement en 4ème position pour le nombre de sites, derrière les sites en anglais, allemand et espagnol.
Nombre d'éditeurs francophones actifs: stable, environ 400 actuellement.
Nombre d'éditeurs actifs répartis dans le monde entier: 10 569 (données au 16 juin 2002)

-> Lamine (http://www.lamine.com), qui gère l'annuaire de MSN France, s'est rapproché de \TEXTUEL, agence de médias de marques et d'entreprises, et a été intégré à TBWA\France (http://www.tbwa.com/). Voir le communiqué :
http://www.textuel.lamine.com/presse/presse.asp?numero=38&annee=2002

-> Pour savoir de quand date la dernière version de votre page sur Google, allez ici :
http://search.cometsystems.com/search.php
le site utilise la base de données de Google. Tapez l'adresse de votre site, par exemple www.abondance.com :
http://search.cometsystems.com/search.php?qry=www.abondance.com&x=28&y=10
Cliquez sur le lien : [Archived copy]. Le message suivant s'affiche, indiquant la date d'indexation de votre site :
The page you see is not a LIVE page of http://www.abondance.com/.
It is an archived copy dated Jun 8, 2002.

Les nouveaux entrants dans l'annuaire des outils de recherche régionaux

Retour au sommaire de la lettre

Pour obtenir tous les sites :
http://annuaire.abondance.com/

-> Correzeweb (région Limousin)
http://www.correzeweb.com/limousin/modules.php?op=modload&name=Web_Links&file=index

-> Ardecheinfo (région Rhône Alpes)
http://www.ardecheinfo.com/

Cherchez, référencez-vous (nouveaux outils ou rappel d'outils existants)

Retour au sommaire de la lettre

-> Poossin
http://www.poossin.com/
Nouvel annuaire francophone.

-> Sovar
http://www.mylinea.com/jbenard/
Annuaire francophone ("slovar" signifie dictionnaire en russe).

-> Big Annuaire
http://www.bigannuaire.free.fr/
Nouvel annuaire francophone

Contenu (sites proposant du contenu ou des fonctions intéressantes)

Retour au sommaire de la lettre

-> Search Engine Marketer
http://www.semlist.com
Nouveau site sur le référencement et le marketing de sites web. Lancement le 16 juillet 2002.

-> Spiders
http://4webhelp.com/spiders/spidersa.shtml
Une nouvelle adresse qui propose les noms (surtout) et numéros IP (parfois) des spiders des principaux moteurs.

Outils (logiciels et sites web proposant des fonctionnalités utiles)

Retour au sommaire de la lettre

-> Google©searchtool v2.5
http://www.frysianfools.com/ggsearch/
Logiciel qui se sert des fonctionnalités de recherche de Google.

-> Whois
http://swhois.net/
Trois outils indispensables lorsqu'on s'interrèsse aux adresse IP des sites web.

-> Yooda
http://www.yooda.com/
Outil de contrôle du positionnement d'un site web sur les outils de recherche.

Revue d'URL

Retour au sommaire de la lettre

-> 131 façons...

http://sewatch.com/searchday/02/sd0711-linktips-long.html
131 façons pour augmenter l'indice de popularité de votre site web...

-> Deep Linking
http://searchenginewatch.com/searchday/02/sd0710-linkalert.html
Quelques informations sur ce qu'il faut savoir au niveau juridique au sujet du "Deep Linking" (mise en place de liens vers des pages internes à certains sites et non pas vers la page d'accueil).

-> Sexe
http://www.uzine.net/article1697.html
Amusant article sur les requêtes ayant trait au sexe sur le Web...

-> Site Banned by Google? Or Was it the Other Way Around?
http://www.traffick.com/article.asp?aID=99
Votre site a disparu de l'index de Google ? Quelques explications...

-> Positionsquatting
http://www.journaldunet.com/juridique/juridique.shtml
Le risque de se voir voler son nom chez Overture et compagnie, par Maître Arnaud Di Meglio.

-> Flash
http://www.academywebspecialists.com/newsletters/0702.html
Comment référencer un site en Flash...

-> Google
http://slashdot.org/article.pl?sid=02/07/03/1352239
Interview de Graig Silverstein, Directeur Technique de Google.

-> Liens promotionnels
http://www.journaldunet.com/dossiers/liens/
Un dossiers complet sur le positionnement publicitaire (Espotting, Overture, etc.)

-> Content is king
http://www.thedmco.com/articles/article19.asp?dtrendsjune
Comment utiliser son contenu pour être mieux visible sur les moteurs de recherche.

-> Image Search and Meta Tags: Copyright Issues?
http://www.clickz.com/search/opt/article.php/1379951
La recherche d'image et les droits d'auteur : un sujet épineux...

-> Métamoteurs
http://www.guerreco.com/sections.php3?op=viewarticle&artid=25
10 métamoteurs utiles dans une recherche d'information...

Merci pour votre lecture... Pour toute suggestion : oa@abondance.com