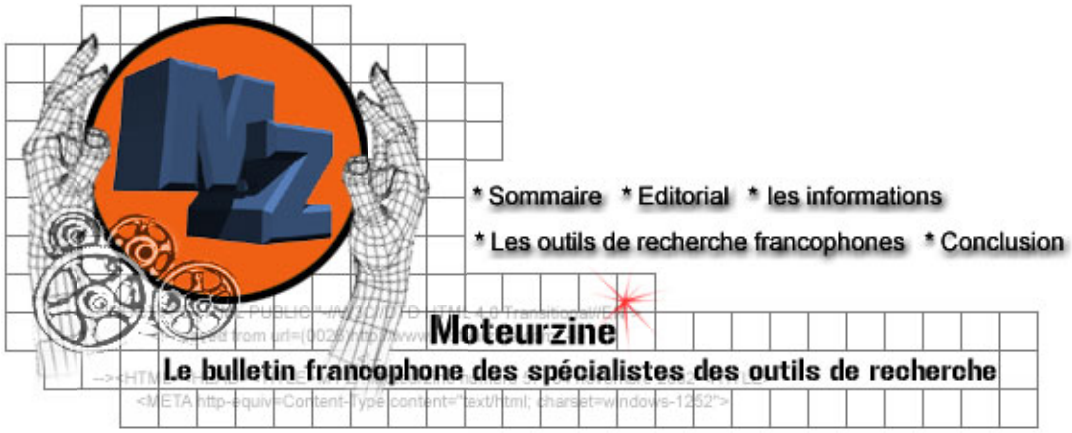
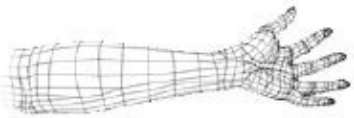
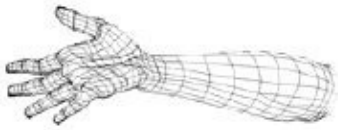


numéro nonante−six (#96) **le mercredi 03 novembre 2003**



Sommaire

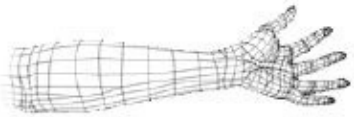


Le dernier Moteurzine en html

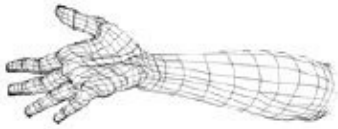
Le dernier Moteurzine en PDF

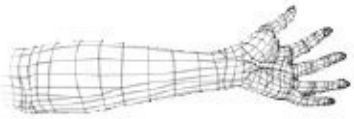
Les archives de Moteurzine

Ecrire à Moteurzine

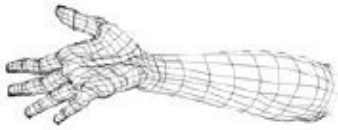


Espace publicitaire





Éditorial



MoteurZine ... mise à jour et passage à la version 5

par CHRis HEDE
Enfin et Moteurzine d'un côté, Développement et Référencement d'un autre.

e−mail
site

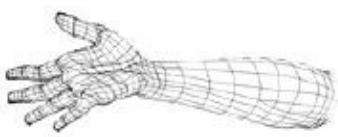
Bonjour à toutes et à tous,

Comme vous pouvez le constater, chaque intervenant sur MoteurZine est désormais présenté à travers une petite fiche. Cela vous permet de mieux connaître ceux qui partagent avec vous leur savoir et vous donnent toujours de bons conseils.

Etant plus que très en retard, je ne m'attarde pas au niveau de l'éditorial et vous laisse lire tranquillement Moteurzine !

Bonne lecture...

Rechercher sur Enfin



Le monde selon Google

Google AdSense : des AdWords sur vos sites

par Olivier Duffez
Editeur de WebRankInfo, il propose aux professionnels des prestations de conseil et référencement sur WebRankExpert.

e−mail
site

Depuis début 2003, Google proposait aux sites anglophones de diffuser des publicités AdSense. Il s'agit de liens publicitaires textuels tels que ceux qui apparaissent sur la droite dans les pages de résultats de Google (les AdWords). L'intérêt majeur de ce système est qu'il détermine de manière automatique le contexte de la page, et choisit en conséquence quelles publicités afficher. Ainsi, les taux de clics sont bien meilleurs quand la pertinence est au rendez−vous. Depuis la semaine dernière (à ma connaissance), les AdSense sont accessibles aux sites non anglophones (pour l'instant : français, espagnol, italien et allemand en plus de l'anglais). J'ai donc testé pour vous ce système sur WebRankInfo et vous livre ici quelques premières impressions.

Différences entre AdWords et AdSense

Le système des Google AdWords est le système publicitaire de Google. Les AdWords sont des emplacements publicitaires dans les pages de résultats de Google, contenant des liens textes vers les annonceurs. Les annonces sont choisies en fonction des mots tapés par l'internaute : si un annonceur a acheté un des mots tapés par l'internaute, alors sa publicité apparaîtra. Cependant, Google limitant à 8 le nombre de liens publicitaires par page, un sytème d'enchères permet de départager les annonceurs ayant acheté les mêmes mots. Le prix à payer par clic dépend de la concurrence : les mots "immobilier" ou "hotel" seront sans doute plus chers que "chocolat" ou "café".

On peut relever quelques différences entre les AdWords et les AdSense :

- les AdWords n'impliquent contractuellement que l'annonceur et Google, tandis que dans le cas des AdSense, Google reverse un pourcentage des gains à l'affilié (celui qui diffuse la publicité sur son site).
- les AdWords sont uniquement sur les sites de Google, alors que les AdSense sont sur n'importe quel site affilié à Google.
- les AdWords "réagissent" aux requêtes des internautes qui font des recherches sur Google, tandis que les AdSense sont choisis automatiquement en fonction du

contenu et du contexte de la page.

Principe de Google AdSense

Après inscription et acceptation du site par l'équipe de Google, on peut commencer à afficher des publicités. Il suffit d'insérer un code JavaScript, identique pour toutes les pages (et même pour tous les sites si vous en avez plusieurs). Les premières heures, vous ne verrez que des publicités pour des sites "humanitaires", qui ne vous rapporteront rien. Mais très rapidement, de nombreux robots de Google vont venir indexer toutes les pages de votre site. Pour savoir facilement et précisément quelles pages ils viennent indexer, et quand, vous n'avez qu'à installer RobotStats, c'est gratuit ! Une fois que les pages ont été indexées et analysées, les publicités seront ciblées en fonction du contenu de chaque page.

Personnalisation des publicités

Sur l'interface AdSense du site de Google, vous pouvez personnaliser les publicités à afficher sur votre site. Pour l'instant elle reste assez sommaire, puisque vous pouvez seulement choisir :

- les couleurs (bordure, fond, titre, texte, lien)
- le format (forme et taille)
- jusqu'à 200 URL de concurrents pour lesquels aucun publicité ne sera affichée

A vous de trouver la meilleure combinaison de couleurs et d'organisation dans la page...

Rémunération

Google interdit aux affiliés de divulguer des informations condidentielles telles que le taux de clic ou la rémunération. Je ne dirai donc rien sur mes premiers résultats... Sachez que le système est du CPC (Coût Par Clic), c'est−à−dire que l'annonceur ne paie pas les affichages mais seulement les clics. Google reverse ensuite une partie à l'affilié. Le fait que la publicité soit très ciblée, et dans un format textuel (au lieu de la traditionnelle bannière animée), permet d'atteindre de bons rendements.

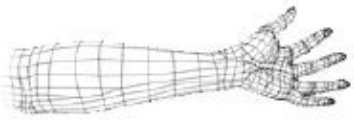
Optimisation des AdSense

De la même manière que des pages peuvent être optimisées pour le référencement, vous pouvez optimiser vos pages pour maximiser vos revenus publicitaires via les AdSense. D'une part, le look et l'emplacement de la publicité doivent être étudiés en fonction du site et du profil des internautes ciblés. D'autre part, il faut savoir que le prix du clic est variable : il dépend de la concurrence sur le ou les mots qui ont été achetés par l'annonceur. Il suffit donc parfois d'utiliser des mots précis dans une page pour faire apparaître des publicités plus rentables... A condition bien sûr, comme pour les optimisations du référencement, de ne pas tromper l'internaute. A quoi bon parler d'immobilier dans un site de recettes de cuisine sous prétexte de générer des meilleurs revenus publicitaires ?

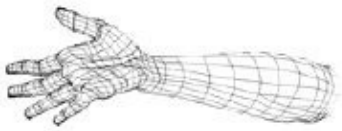
Pour en savoir plus

Tout d'abord, allez lire en détails les informations et les conseils donnés par Google sur

www.google.com/adsense. Ensuite, venez découvrir et partager vos expériences sur le forum AdSense de WebRankInfo.



Les entreprises et les outils de recherche



Enquête sur les idées reçues dans le métier des outils de recherche

e−mail
site

par Jean−Claude Benard
de Quadramultimédia.

Dans notre monde de certitudes et de faits établis, s'il est une industrie qui s'illustre, c'est bien celle qui tourne autour des outils de recherche.

Appartenant moi−même à cet environnement, je suis confronté quotidiennement à des remarques ou réflexions de mes clients entreprises.

Flaubert écrivit à son époque un dictionnaire des idées reçues. En forme d'hommage à celui−ci nous avons décidé de créer polémique et réflexion autour d'un thème qui nous est cher : Les outils de recherche et les acteurs qui vivent dans cet environnement.

Nous sommes tous, par habitude ou insouciance amenés à véhiculer des lieux communs ou idées reçues. Il m'a semblé intéressant de donner la parole à des entreprises sur ce thème.

Au cours des prochaines chroniques, nous interrogerons des chefs d'entreprises appartenant aux autres industries pour leur soumettre des propos ou idées reçues qui ont la vie dure dans notre métier et connaître leur vision et réponses.

Pour commencer cette série, nous avons décidé de commencer nous−mêmes par quelques phrases souvent entendues et qui tiennent lieu d'exemples

Un site Web bien référencé augmente les ventes des entreprises

Malheureusement, être bien référencé n'est pas gagner.

Pourtant beaucoup de confrères que nous avons croisé au détour de congrès ou réunions légèrement nombrilistes, les internautes sont une sorte de troupeau de moutons martiens virtuels.

L'internaute est un consommateur de plus en plus avertit qui profite du Web pour s'informer. Nous avons pu constater auprès de nombreux référencés que les cycles de décisions sont

d'autant plus long que l'internaute s'abrite derrière une adresse électronique.

De plus, un mauvais produit présenté sur un site Web ne gagnera pas en qualité parce que son référencement est particulièrement bien fait et les problèmes de déficits de contenu ou de produits obsolètes ne peuvent être résolus par un bon algorithme

Se persuader et faire croire à un client que ses « rossignols » pourront être mieux vendus parce qu'il est présent dans les deux ou trois premiers résultats de recherche n'honore pas notre métier et n'a aucun effet auprès des internautes.

Référencer n'est pas seulement un acte informatique mais avant tout un travail d'accompagnement des entreprises qui souhaitent obtenir un bon positionnement sur le Web.

Un audit du site du client suivi des adaptations nécessaires fait partie de notre travail. Si le client refuse toute évolution, il est préférable qu'il assume seul ses déboires.

C'est à cette seule condition que nous pourrons affirmer qu'un bon référencement PEUT aider à vendre

Les liens sponsorisés coûtent cher

Prenons pour exemple le cas d'un cabinet de courtage en assurance.

La majorité de ceux−ci distribuent des contrats concernant les personnes (santé, prévoyance, décès) les biens (immobilier, véhicules) et pour certains la responsabilité professionnelle.

Nous avons pu constater que sans avoir eu un entretien précis avec le client sur ses motivations et intentions, le montant que celui−ci va engager pourra varier du simple au démentiel.

Liste de mots clés génériques

Recherches sur le Web effectuée à partir des statistiques de requêtes ciblées des internautes sur les outils de recherche partenaires de ESPOTTING au mois de novembre.

Mots	Nombre de requêtes prévues	Prix de la première place	Coût estimatif en €
Assureurs	1519	0,21	318,99
Assurance vie	385094	0,98	377392,12
Assurance habitation	4511	0,25	1127,75
Assurance santé	12549	0,40	5019,60

Total théorique			383858,46

C'est vrai que si on ne tient pas compte de la structure juridique du courtier et de son approche habituelle de commercialisation, nous ne donnons pas cher des 200 ou 300 ? qui seront engagés dans sa campagne pour un résultat qui s'annonce décevant

Liste de mots plus adaptés et ciblés

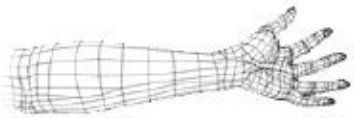
En étant cohérent, notre cabinet de courtage pourrait obtenir un trafic beaucoup plus qualifié en utilisant des mots plus précis

Mots	Nombre de requêtes prévues	Prix de la première place	Coût estimatif en €
Courtiers assurances	48	0,14	6,72
Cabinet courtier assurances	27	0,10	2,70
Assurance Paris	22	0,11	2,42
Assurance autos Paris	27	0,40	10,80
Devis assurance autos	818	0,10	81,10
Devis assurance santé	72	0,16	11,52

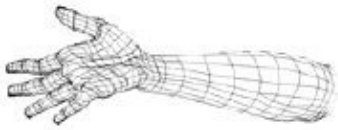
Total théorique			115,96

Cet exemple met en lumière un seul mot : l'optimisation. Essayez de le reproduire avec d'autres thèmes, vous n'êtes pas au bout de vos surprises.

Suite à la prochaine rubrique...



Le référencement marketing



Optimiser le marketing de votre balise Description (1)

par Gautier Girard
Consultant en création de valeur et de trafic ciblé sur internet.
e−mail
site

Maintenant que nous sommes en mesure de cerner le bon compromis entre référencement et optimisation marketing du référencement, arrêtons nous quelques instants sur l'optimisation de la balise META description.

Cet article a pour but d'optimiser son affichage dans les pages de résultats tout en conservant, bien entendu, une excellente qualité d'optimisation en matière de référencement.

Nous commençons aujourd'hui avec l'intérêt de l'optimisation de la balise description.

Quel est l'intérêt d'optimiser cette balise ?

La balise meta description est importante pour deux raisons :

1− Elle est utilisée par les moteurs de recherche, à la fois dans l'indexation, à la fois dans

l'affichage en pages de résultats sur certains.

2− Une "bonne" balise description peut−être reprise telle quelle par des netsurfers intéressés par l'ajout de votre site dans un annuaire.

Elle peut également être utilisée par des webmasters souhaitant faire un lien vers votre site pour le prendre comme référence. Quoi de plus simple que de copier/coller une balise description et de la placer telle quelle sur son site ?

Que ce soit pour les annuaires ou des sites lambda, cela arrive plus fréquemment qu'on ne l'imagine. Et c'est tout à fait normal, l'humain est paresseux ;−).

1− Moteurs de recherche

Certains moteurs de recherche, comme Alltheweb ou Dir.com, affichent dès la page de résultats le contenu de la balise description. Voici un exemple ci−dessous pour une requête +gîte +alpes:

En plus du contenu affiché par le moteur lui−même, vous disposez d'un espace supplémentaire pour convaincre le visiteur qu'il doit cliquer sur VOTRE site, et non pas celui se trouvant au dessus ou en dessous.

2− Annuaires et sites

Même si les annuaires de recherche tendent depuis longtemps vers les référencements "express" (payants), le besoin en contenu de qualité est toutefois toujours d'actualité.

Il arrive par le plus heureux des hasards qu'un netsurfeur tombe sur votre site et décide de l'ajouter gratuitement dans la catégorie qu'il gère. Si votre balise description correspond aux critères d'acceptation sans besoin d'être retouchée, alors c'est tout bon : le netsurfeur n'a plus

qu'à copier/coller une description que vous contrôlez.

Pour des sites lambda, cela arrive fréquemment aussi. Quoi de plus normal pour un webmaster que d'ajouter des sites jugés pertinents pour compléter son contenu ?

Pour débuter, des principes de base

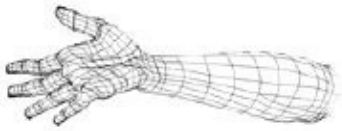
Quelques principes simples permettent d'optimiser la balise meta description. Nous allons étudier cet aspect dans les prochains numéros de MoteurZine.

Il s'agit de répondre à 3 questions principales :

− La description répond−elle au besoin formulé par l'internaute dans sa requête ?
− La description donne t−elle envie de cliquer sur ce lien ?
− La description est−elle optimisée pour le référencement ?

Dans le prochain MoteurZine nous verrons comment répondre au besoin formulé par l'internaute. Et parce que les vacances d'hiver approchent à grands pas, je vous donnerai un peu plus envie de prendre un gîte dans les Alpes ;−).

Humour et Internet



Voyeurisme sur Internet

par Frédéric Lepage
« Un crayon, un ordinateur et des idées... Plein d'idées ! »

e−mail
site

Référencement de A à Z

Référencement dans les moteurs de recherche

par Marie Pourreyron
Consultante indépendante en référencement et promotion de site internet. Gérante d'Altitude Référencement.
e−mail
site

Nous avons vu la procédure à respecter pour s'inscrire dans les annuaires généralistes ou thématiques.

Nous allons voir aujourd'hui les moteurs de recherche.

Un moteur de recherche est différent d'un annuaire car fonctionne au moyen de robots (les crawlers ou spiders) qui naviguent de pages en pages en suivant les liens et enregistrent au fur et à mesures les pages visités. Lorsqu'une url est stockée dans l'index d'un moteur, elle est revisitée à fréquence variable (selon le moteur) et l'index est remis à jour périodiquement afin de refléter les changements des pages web.

Les moteurs de recherche sont nombreux sur Internet mais une minorité d'entre eux assurent la majorité du trafic.

Le moteur de recherche le plus connu et utilisé de nos jours est sans conteste Google qui génère environ 60% du trafic issu des outils de recherche. Cela veut−il dire qu'il ne faut se référencer que sur Google ? Certes non, ce serait une erreur car vous perdriez en visibilité.

Consultez un baromètre du référencement pour en savoir plus :

- Baromètre Abondance
- Baromètre Adoc
- Baromètre Indicateur

Comment soumettre son site à un moteur de recherche?

La soumission est simple et pas forcément indispensable. En effet si votre site a déjà des liens d'autres sites pointant vers lui, la plupart du temps il ne sera pas toujours nécessaire de soumettre aux moteurs de recherche, les robots d'indexation suivant les liens présents dans les pages, ils trouveront votre site.

Pour soumettre votre site à un moteur de recherche c'est très simple, il suffit de se rendre sur le moteur en question, de trouver le lien "suggérer un site" et de soumettre votre url, éventuellement un titre et un email.

Autre possibilité, utilisez un outil comme celui proposé par Yooda qui vous permettra de gagner du temps et de vérifier la présence de votre site dans l'index du moteur.

Comment savoir si les robots d'indexation sont passés sur mon site ?

Il suffit de regarder dans vos fichiers journaux. Chaque moteur de recherche possède un crawler spécifique (googlebot pour Google par exemple)

Pour connaître la correspondance entre le nom d'un crawler et le moteur de recherche pour qui il "travaille" vous pouvez consulter ce site !

Combien de temps dois−je attendre avant d'être présent dans les moteurs de recherche ?

Les délais sont extrêmement variables, ils peuvent aller d'une semaine à plus de 6 mois selon les moteurs.

Est−ce payant?

Les moteurs de recherche sont majoritairement gratuits mais certains proposent une indexation payante, permettant d'accélérer le processus et de faire rafraîchir ses pages par le moteur de façon régulière.

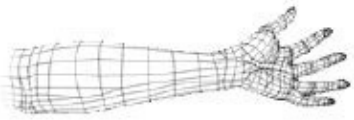
Pourquoi me demande−t−on de recopier un code pour pouvoir inscrire mon site ?

Afin de limiter les inscriptions massives provenant de logiciels de référencement, les moteurs obligent à recopier un code fourni sous forme d'image, ce que ne peut pas faire un logiciel.

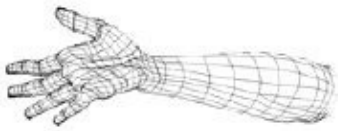
L'utilisation d'un logiciel pour soumettre mon site est−il indispensable ?

Non, grace à la question précédente, vous devez avoir compris pourquoi! Par contre ne confondez pas logiciel de soumission et applicatif aidant à la soumission tels que ceux proosés par Yooda et Refpayant ;−)

La prochaine fois nous verrons les accords passés entre les portails, les annuaires et les moteurs : qui fournit quoi à qui?



Optimisation et indexation commerciale



Brasser des affaires sur le net, avec 10 détectives virtuels

par Pierre Frigon
Auteur et l'éditeur des Histoires à succès de l'internet : L'encyclopédie Mondiale du Commerce Électronique Interactif et Personnalisé.

e−mail
site

LES AJUSTEMENTS POST−GUTENBERG

Gutenberg s'en était aperçu. Lancer un nouveau média de communication n'est pas qu'une mince affaire. Il finit ses jours complètement ruiné et abandonné de tous. Même, qu'un peu plus tard dans l'histoire de son invention, certains en étaient venus à brûler certaines ?uvres publiées qui ne faisait pas quorum avec les valeurs et les pouvoirs de l'époque. Même la plus merveilleuse des nouveautés trouve ses récalcitrants et ses détracteurs.

Que ce soit l'avènement de la radio, de la télé ou de l'Internet, adapter les pratiques courantes d'affaires aux nouveaux moyens de communication demande toujours... quelques ajustements.

LA PROFONDE SIMPLICITÉ DES SURFACES

À la surface des choses, tout peut paraître si simple, comme beaucoup d'entreprises l'estiment à la lumière des sites Web dont ils se dotent. Après tout, c'est aussi simple que d'imprimer une page de la bible de Gutenberg. Un rouleau passe, ou une souris clique et le tour est joué... On existe. On est publié. Là s'arrête la réalité d'affaires d'encore trop d'entreprises... alors que les pertes fiscales de sites Web d'amateurs, elles, continuent.

À ce titre, même un Boeing 747 voguant à 30 000 pieds parait simple, pour le simpliste. Mais, le véritable problème survient quand on cherche à devenir rentable, avec son site Web. Autrement, aussi détaché de la réalité que soit ce Jet dans les nuages, le simpliste y comprendra toujours que... plus c'est loin, plus c'est simple.

L'EXEMPLE DES DÉSÉQUILIBRES BUDGÉTAIRES DES PUBLICITÉS WEB

Dans certaines de mes récentes rubriques, je mentionnais le déséquilibre flagrant entre les coûts des publicités payés par les entreprises pour l'achat de mots clés à l'enchère, dans les engins de recherche, versus l'optimisation scientifique et plus permanente des pages Web. Cette variation astronomique des coûts devient évidente pour qui s'approche d'un peu plus près de cette équation.

Il en va de même pour la gestion compétitive de son site Web. Toute entreprise qui possède des compétiteurs à proximité, aura l'occasion de se tenir au courant des tentatives d'accroissement des parts de marché de ceux−ci. Cette information l'incitera à contrer cet effort de surclassement en y allant de sa stratégie propre, avant qu'il ne soit trop tard.

PREMIÈRE VICTOIRE OU DERNIÈRE DÉFAITE

Mais sur le Web, là où ce sont des centaines et des milliers de compétiteurs qui s'enfilent et s'affichent avant et après votre entreprise dans les résultats de recherches, à tout coup, la chose devient plus corsée. L'Internet est le lieu où doit s'appliquer une vigilance hors pair, quant à l'intelligence compétitive et d'affaires. Et la bataille se joue sur plusieurs fronts.

Le premier terrain de cette bataille se joue dans la justesse et le nombre des mots clés qui définissent vos activités commerciales et celles de ces compétiteurs. Être vu est la première victoire ou la dernière défaite. À regarder ça de près, aujourd'hui, vos compétiteurs pourraient vous laisser l'avenue complètement libre, tant les meilleures pratiques d'affaires Internet sont ignorées par les décideurs aussi bien que par leurs webmestres. Par exemple, comprennent−ils l'impact financier des nombreux textes encore livrés en Flash ou en JPG, ou celui d'un nom de domaine illisible, parce que pris en un bloc, pour les engins automatisés de leur rentabilité financière?

Les 19 corps de métiers nécessaires pour créer et opérer un site Web rentable ne sont pas encore captés dans le *radar* de la majorité des entreprises... et c'est mieux ainsi pour ceux qui regardent ces choses de plus près. À eux d'en profiter.

UN JEU VIDÉO POUR LES PETITS ET LES GRANDS

Cette joute ressemble aux celle des jeux vidéo, où plusieurs niveaux existent, mais où seulement les plus futés savent comment s'y rendre. Quant aux débutants ou aux amateurs, ils n'en soupçonnent même pas l'existence, sinon dans dix ans!

Les plus futés font régulièrement leurs recherches pour comprendre les mouvements compétitifs de leurs adversaires sur le Net. Si anciennement une simple promenade aux tablettes de son compétiteur local offrant les mêmes produits que ceux de son entreprise était chose courante et facile, faire le tour du Web pour voir ces choses est beaucoup plus difficile aujourd'hui. Les sites Web naissent comme les champignons. À nous de découvrir ceux qui ont la capacité d'empoisonner notre rentabilité.

C'est pourquoi ceux qui font de l'Intelligence Compétitive ou de la Veille Concurrentielle possèdent l'expertise intellectuelle et matérielle d'effectuer cette démarche régulièrement. Puisque l'Intelligence Compétitive est un domaine en croissance fulgurante, à cause de la mondialisation de l'Internet, plusieurs entreprises offrent de pareils services.

LE DÉVELOPPEMENT VS LA DISPERSION

Pour suivre de près les nouveaux sites Internet, l'utilisation du moteur de recherche francophone http://www.kartoo.com peut s'avérer plus que bénéfique, puisqu'il permet l'abonnement "Push" aux alertes spéciales qui identifient tout nouveau site Web dans le secteur de son choix. Pour aller plus loin, d'autres sites, services Web ou logiciels permettent un suivi très serré des activités de ses compétiteurs.

Néanmoins, effectuer de telles recherches demande du temps et des compétences pointues d'affaires, évidemment. Connaître l'Internet n'est pas suffisant. Il faut aussi le comprendre dans ses mécanismes secrets et intimes. C'est pourquoi plusieurs considèrent désastreux le développement de nouvelles compétences complexes à maîtriser... s'ils ne peuvent continuer à s'occuper à fond de leurs véritables affaires principales.

Une solution d'impartition est souvent la meilleure voie à suivre − et de loin, car elle permet la garantie de la pertinence des résultats sur lesquels l'avenir de l'entreprise repose, et l'économie de ne pas assumer la permanence d'un poste interne supplémentaire, en acquisition lente, progressive et inexpérimentée des résultats guidant la barque de la compétitivité corporative.

10 DÉTECTIVES VIRTUELS

Pour se mettre l'eau à la bouche, ou pour regarder tout ça de plus près et jauger pour soi la complexité de la Veille Concurrentielle, voici une liste de 10 "détectives virtuels", aptes à vous permettre le tour de votre quartier ou de la planète, en un clic − si vous en avez l'appétit!

Panorama des principales solutions de veille

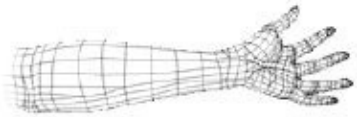
Editeur	Solution	Commentaires
Aignes	**WebSite−Watcher**	Contrôle si des mises à jour ont été faites sur plus de 100 sites par minute et télécharge les modifications sur le disque dur. Mise en évidence des changements automatiques. Peut passer outre des formulaires en mode POST. Envoi des alertes par courriel ou ouvre le site concerné en cas de mise à jour détectée.
Albert	**AMI Market Intelligence**	Automatise la collecte et l'analyse d'informations situées sur des sites concurrents, des fichiers, des groupes de discussion, etc. Interface Web.
Arisem	**KM Server / Competitive Intelligence**	Surveillance du Web en continu, classification de l'information selon des catégories prédéfinies et sous forme d'arborescence, diffusion de l'information en mode alerte. La KM Server propose, en outre, de multiples fonctionnalités de travail coopératif.
BEA Conseil	**KB Crawl 2.0**	Surveillance de sites Web statiques ou dynamiques (PHP, ASP), nécessitant éventuellement une authentification (identifiant/mot de passe) ou sécurisés par une connexion HTTPS ou SSL. Surveillance de formulaires et de bases de données également. Alertes sélectives par courriel lors de changement de contenu, apparition de mots−clés ou de nouvelle page. Archivage des documents trouvés.
Copernic	**Copernic Agent Professional**	Accès à plus de 1 000 moteurs de recherche répartis en 120 catégories. Surveillance des contenus d'un nombre illimité de pages Web. Résumés des pages Web trouvées et extraction de leurs concepts clés. Alertes par

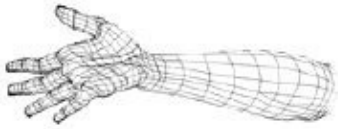
		courriel avec copie de la page où les changements sont surlignés. Intégration dans Internet Explorer et Office 2000/XP par l'ajout d'une barre d'outils dédiée.
Digimind	**Evolution**	Solution intégrée composée d'une plate−forme de base sur laquelle viennent se greffer des modules additionnels. La surveillance de pages ou de sites Web est possible, ainsi que des alertes par courriel en fonction d'un niveau de modification prédéterminé : modification de plus de x % du contenu d'une page, modification des images, des liens par exemple. Le Web invisible, les listes de discussion et les groupes de discussion Usenet sont également surveillables.
Intelliseek	**Marketing intelligence**	Propose une gamme de solutions permettant de scruter l'information disponible sur le Web, à hauteur de cinq millions de pages par jour. L'information peut concerner la marque de l'entreprise, l'impact d'une campagne de publicité, les avis de consommateurs postés sur des forums, etc.
Sinequa	**Intuition / iInternet**	Avec le produit iInternet, le moteur de recherche Intuition indexe pages et sites Internet. Avec le produit iPush, les utilisateurs sont prévenus des résultats de leurs filtrages sélectifs.
Verity	**Verity K2 Enterprise**	Solution intégrée de recherche et de catégorisation, K2 Entreprise indexe de multiples sources de données textuelles (courriels, bases de données, sites Web) et référence automatiquement les nouveaux documents. La solution fusionne et catégorise les résultats issus de recherches sur les index de sources d'information Internet telles que Altavista, Factiva, Google, Hoover ou Moreover.
Wysigot	**Wysigot**	Surveille tout site Web, alerte en cas de nouveaux documents disponibles qui peuvent ensuite être capturés et stockés.

Je remercie vivement Fabrice Deblock et www.journaldunet.com pour ce tableau.

Je suis toujours heureux de répondre personnellement aux correspondants.

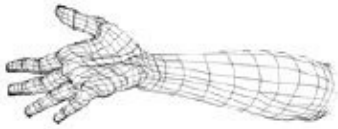
Au plaisir





Vous ICI ???

Enfin, les informations



Vous êtes le responsable d'un moteur ou d'un annuaire ? Vous travaillez dans le référencement ? Communiquez et insérez gratuitement votre actualité ici.

AUJOURD'HUI

ou le monde impitoyable du référencement et des outils de recherche

29/11/2003 : "Légalisé dans quelques jours aux Etats−Unis, le spam pourrait exploser en 2004"

« George W. Bush devrait parapher prochainement une loi permettant aux entreprises d'envoyer des e−mails à caractère commercial tant que les destinataires ne s'y opposent pas. Les efforts judiciaires européens et suisses risquent d'être ainsi anéantis étant donné que la majorité des spammeurs sont américains et qu'il sera très difficile de les poursuivre à l'avenir. »

28/11/2003 : "Les weblogs, nouvelle arme du porno ?"

« Des webloggers américains ont découvert la dernière technique employé par l'industrie du porno pour augmenter le trafic sur leur site : se cacher dans de faux weblogs afin d'être mieux classés dans Google. (...) »

28/11/2003 : "Le nombre d'internautes a augmenté de 25% en France depuis octobre 2002"

« Le nombre d'internautes âgés de 11 ans et plus était de 21,67 millions en France en octobre 2003, soit une progression de 25% par rapport à la population des internautes d'octobre 2002, selon une enquête de Médiamétrie publiée jeudi. »

27/11/2003 : "Pool Entreprises proposent une nouvelle offre attractive de référencement"

Après avoir offert aux sociétés 100 référencements gratuits dans sa base, Pool Entreprises proposent à celles qui n'ont pas eu le temps de soumettre leur site, un référencement à 50% du prix normal. Cette offre est valable jusqu'au 15 décembre 2003 et uniquement pour les 500 premières soumissions.

27/11/2003 : "LiensFr, un outil de recherche francophone"

« LiensFR est un outil de recherche recensant les sites Internet francophones et proposant une recherche soit par mots clés, soit par un classement thématique. »

27/11/2003 : "Les pages Web ont une durée de vie limitée..."

« Si Internet est pour bon nombre de chercheurs un facilitateur en cela qu'il simplifie grandement la recherche d'information et leur évite souvent de passer de longues heures à fureter dans les rayons des bibliothèques universitaires, il se révèle aussi être un casse−tête. Deux scientifiques spécialisés en dermatologie en ont fait l'expérience au cours d'une de leur étude, relate le Washington Post dans son édition de lundi 24 novembre. En voulant relayer dans leurs notes de bas de pages et dans leur bibliographie des adresses de sites Internet ou de pages utiles pour la compréhension de leurs travaux, ils ont pu percevoir à quel point le Web est

éphémère. La rédaction de cette étude ayant nécessité deux ans de labeur, au terme de ce travail, la plupart des liens cités étaient devenus erronés : soit les sites avaient disparu, soit la localisation des pages avait changé. » (...)

26/11/2003 : "Overture revendique 2.000 annonceurs en France"

« Les liens sponsorisés sont−ils en train de démocratiser la publicité ? Pionnier des liens promotionnels (sponsorisés ou contextuels), l'américain Overture, désormais filiale de Yahoo, annonce avoir séduit plus de 2000 annonceurs, dont une forte proportion de PME. »

25/11/2003 : "Espotting Media et AutoPlus annoncent leur partenariat"

« Le spécialiste du lien promotionnel continue d'agrandir son réseau de distribution, et, après avoir annoncé son partenariat avec les différents sites du groupe Emap, intègre dorénavant son annuaire sur le site www.autoplus.fr. (...)
Ce nouvel affilié s'inscrit totalement dans l'esprit du réseau de distribution Espotting, AutoPlus étant un généraliste automobile dont la vocation est de s'adresser à tous les automobilistes français. Il accompagne les lecteurs dans tous les moments de leur vie de conducteur (achat, ventes de véhicules, droit, sécurité, entretien...). Son approche est pratique et accessible. C'est un journal au service de ses lecteurs, à la fois descriptif et prescripteur, tout comme Espotting. (...) »

25/11/2003 : "Ouverture de l'Annuaire des voyages"

L'annuaire des voyages est un outil de recherche 100% dédié aux voyages. Sur cet annuaire, il n'y a pas de bannalités et ni de liens publicitaires. C'est uniquement un annuaire complet sur les voyages et ces problématiques.

25/11/2003 : "Le monde selon Google"

« Une commodité presque banalisée ; un tuyau de plus, souvent branché par les mêmes sociétés qui apportent eau, téléphone ou télévision : la Toile d'Internet ne suscite plus ni grandes angoisses ni envolées lyriques. Tout est là, à portée de souris, dans les mémoires gigantesques des moteurs de recherche... Mais quels sont leurs critères de choix ? Quels biais introduisent−ils dans les représentations ? Dissection du premier d'entre eux, Google. »

24/11/2003 : "Ouverture d'un annuaire dédié au tourisme au Canada"

Les français sont passionnés par le Canada. Ils aiment visiter ce pays qu'explora, entre autres, Jacques Cartier. Pour partir en vacances dans une des provinces du Canada ou bien en voyage d'affaires, visitez l'annuaire du Canada qui vous aide à vous organiser pour ce voyage.

22/11/2003 : "Overture lance un outil de tracking publicitaire payant"

« Overture, l'un des principaux acteurs du référencement payant, a lancé sa "Console marketing", un logiciel de mesure de la performance des campagnes en ligne. Il entend ainsi répondre à un besoin essentiel des marketers (mesurer les résultats que permettent les nouvelles techniques) et justifier, si possible, le recours à ses services. Il est vrai qu'entre l'e−mail marketing, les liens promotionnels et l'e−pub, l'arsenal du marketing en ligne s'est non seulement étendu mais aussi beaucoup complexifié. Par la même occasion, les techniques sont de plus en plus spécifiques, et de nouvelles disciplines se sont imposées d'elles−mêmes, comme le "search engine marketing". (...) »

21/11/2003 : "Yahoo rachète un moteur de recherche chinois pour 120 millions de dollars"

« La société américaine Yahoo Inc. a annoncé qu'elle allait débourser 120 millions de dollars pour acheter la 3721 Network Software Company, propriétaire de la technologie du moteur de recherche en langue chinoise 3721.com.
L'acquisition doit s'étaler sur une période de deux ans et débuter au premier trimestre

2004.
Grâce aux programmes mis au point par 3721 NSC, les utilisateurs chinois de l'internet peuvent, au lieu de taper des noms de domaine ou des adresses internet en anglais, saisir directement des caractères chinois.
L'internet compte près de 70 millions d'utilisateurs en Chine et ce nombre augmente rapidement. (...) »

20/11/2003 : "ITV choisit Espotting tandis que 01net, Boursier et 20minutes optent pour Overture"

(...) « Après MSN, Wanadoo, Tiscali ou encore Yahoo, Overture délivre désormais ses liens contextuels sur les sites 01net.com, 20minutes.fr et Boursier.com. Avec les liens contextuels, les résultats d'Overture sont affichés sur les pages des sites de ces partenaires en fonction de leur contenu éditorial. Grâce à la signature de ces accords locaux qui viennent compléter le renouvellement de l'accord mondial avec MSN, Overture revendique désormais le réseau le plus puissant et le plus efficace du marché des liens contextuels en France.
De l'autre côté de la manche, Espotting, qui revendique le leadership sur le continent européen du positionnement payant à la performance, remplace justement son concurrent Overture sur tout le réseau d'ITV et intègrera ainsi son moteur de recherche et ses liens contextuels sur tous les différents sites internet appartenant à ITV (ITV Football, Formula 1, Pop Idol, Qui veut gagner des millions, etc). »

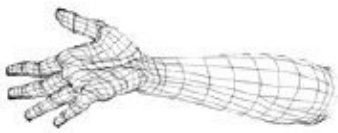
LES ARCHIVES

ou il y a un an, l'actualité du moment

26/11/2002 : "Et non, le domaine google.no n'appartient toujours pas à Google !"

Google a porté devant la justice norvégienne le problème de squatting de son domaine google.no. La justice norvégienne a envoyé paître Google et a donné raison à SMSfun, la société possedant le domaine google.no. Il va de soit que Google va faire appel.

Les outils de recherche francophones



29 novembre 2003
Echange de liens
L'annuaire gratuit pour boostez votre site.

Adulte gratuit
Annuaire de sexe gratuit propre et simple à utiliser.

27 novembre 2003
Officiel des iles
Portail des îles du monde.

26 novembre 2003
Annuairesexe.biz
Annuaire du sexe gratuit avec un classement original basé sur la qualité et la popularité.

25 novembre 2003

Annuaire des voyages
Annuaire dédié aux voyages.

24 novembre 2003
X−recherche
une page pour interroger plusieurs sites.

JV1
Outil de recherche de sites, d'informations et de nouvelles sur les jeux vidéo.

23 novembre 2003
Provence touristique
Annuaire cartographique, thématique et par mot clés dans la Provence.

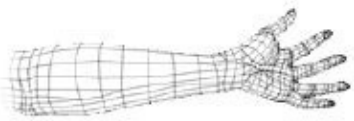
Queyras touristique
Annuaire cartographique, thématique et par mot clés dans le Queyras.

22 novembre 2003
Annuaires Afrique
Annuaires officiels de 20 pays d'afrique en Francais

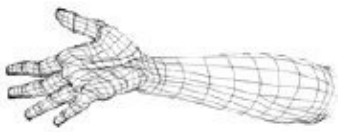
21 novembre 2003
Allez UK
Moteur spécialisé dans l'industrie du tourisme pour le marche anglais.

The scout annuaire
Annuaire thématique sur le scoutisme.

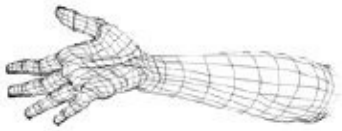
20 novembre 2003
Briançon−Vauban
Annuaire des sites situés à Briançon ou le briançonnais dans les Hautes Alpes.



Auto Promotion



L'article / l'entretien



Les techniques de crawl évoluées : le problème de la fraîcheur des index

Philippe YONNET
e−mail
site

par Philippe Yonnet
Directeur du Département Informatique et Internet de la SA
STUDYRAMA, et administrateur du Webmaster−Hub.

Les techniques de crawl ont beaucoup évolué en quelques années. Et de nombreux moteurs ont adoptés des technologies nouvelles, pas forcément très connues des webmasters. Ceux qui consultent leurs logs régulièrement ont pu remarquer des changements subtils dans le comportement de certains robots (Googlebot notamment). Nous allons vous en dévoiler les raisons, et les nouveaux objectifs assignés à ces crawlers.

Indexer vite, indexer tout, mais indexer poliment

La taille de la toile ne cesse d'augmenter. Cette croissance touche à la fois le nombre de pages, mais aussi la taille moyenne de chaque page. Par ailleurs, la part des pages de texte diminuent au profit des pages contenant des images, des videos, des animations flash etc...

Dans le même temps, la bande passante disponible pour crawler le web est toujours limitée. Il est donc indispensable, et même stratégique, d'optimiser les crawlers pour qu'ils puissent indexer la Toile de manière efficace.*

Cette efficacité se mesure en fonction de trois facteurs :

1°) L'indexation doit être rapide : c'est la condition sine qua non pour qu'un robot puisse passer souvent et assurer que les pages figurant dans l'index aient une "fraîcheur suffisante"

2°) L'indexation doit être complète : on sait qu'elle ne peut pas être exhaustive, car certaines portions du web ne sont pas reliées entre elles par des liens. Mais le robot doit être capable d'indexer une portion significative de la Toile

3°) Dans le même temps, l'indexation doit respecter les sites visités : notamment, le robot doit tenir compte des instructions du fichier robots.txt, et ne pas aspirer des centaines de pages à la seconde...**

Les enjeux d'une indexation efficace

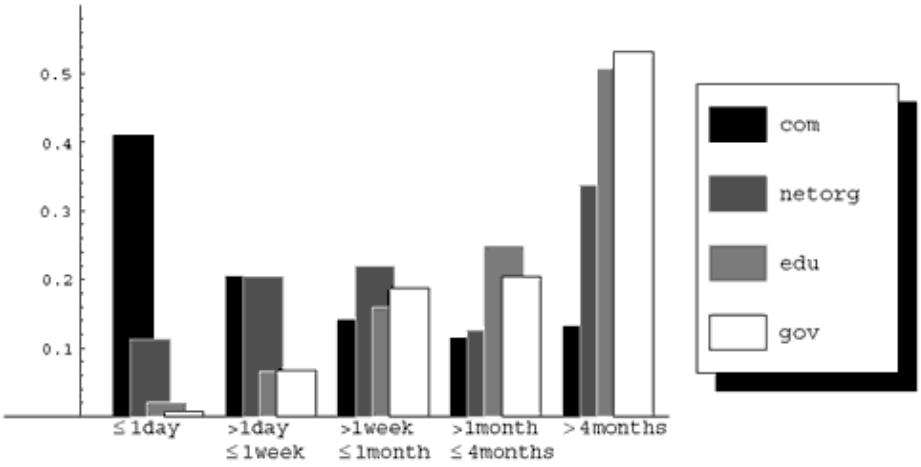
Pour développer un moteur de recherche encore plus pertinent, il convient d'améliorer les techniques d'indexation dans trois domaines différents.

- D'abord, savoir indexer toujours plus de formats de fichiers différents.
- Ensuite, indexer plus de pages, plus vite, et de manière opportune pour conserver la fraicheur de l'index.
- Enfin, savoir indexer plus de pages dynamiques.

Le problème de l'indexation des formats de fichiers non textuels ou exotiques, et celui des pages dynamiques ne seront pas abordés ici. Ils nécessitent chacun un développement complet.

La fraîcheur de l'index quant à elle représente un enjeu important. Il était mal géré par la plupart des moteurs de recherche grand public jusqu'à une époque récente : et encore aujourd'hui, il n'est pas rare de se voir renvoyer une erreur 404 en cliquant sur une page de résultats. Or les pages "d'actualité" se multiplient sur la Toile, et le rythme de changement des pages ne cesse d'augmenter.

Dans le même temps, le rythme de "mise à jour" varie dans des proportions extrêmes d'un site à l'autre, et même souvent entre les pages d'un même site. Imaginer des stratégies de crawl qui tiennent compte de ces facteurs constitue donc une nécessité si l'on ne veut pas, soit crawler souvent des pages qui ne changent pas, et consommer des ressources utiles, soit crawler de manière trop espacée et rater des changements intermédiaires...



Fréquence de changement des pages en fonction des domaines − extrait de [1]

Les méthodes traditionnelles de crawl : le batch crawling

La méthode de crawl la plus communément utilisée jusque là était l'"indexation par lot" ("batch crawling" en anglais.

Le batch crawling se déroule en trois étapes.

D'abord, on détermine une série d'urls de départ à crawler (les " seed url "). La taille de ce fichier d'urls et la manière de les choisir a d'ailleurs une influence certaine sur le résultat final. Il suffit de penser à ce qui peut se passer si l'on ne choisit que les adresses de site ne figurant que dans une seule catégorie DMOZ par exemple : l'index qui sera créé risque d'être sérieusement "coloré" par ce choix, et des pans entiers du web ne seront pas indexés.

Ensuite, on lance les robots d'indexation qui vont "aspirer" les pages, tout en récupérant les liens qu'elles contiennent vers d'autres pages.

Ces nouvelles url sont ajoutées à une "file d'attente" des liens qui restent à explorer. Lorsque cette file d'attente ne contient plus de nouvelles url, le processus s'arrête. Le crawl est terminé. Et l'on considère que la Toile entière a été indexée.

Les urls ainsi recueillies pendant le crawl serviront par ailleurs d'urls de départ au prochain cycle de crawl...

Ce processus de "batch crawling" décrit en fait ce que les webmasters observaient sur leur site lors des full crawls de Google jusqu'au printemps 2003. La version "deep crawl" de Googlebot était un "batch crawler" typique.

Les inconvénients du batch crawling

L'indexation par lots a de graves inconvénients, dès que l'on veut développer un moteur de recherche "grand public" comportant un très grand nombre de pages. D'abord les crawlers indexent l'ensemble des pages à chaque cycle, y compris celles qui ne changent jamais. Le processus est donc long, il dure dans la pratique plusieurs jours, jusqu'à une semaine complète. Or, c'est seulement à la fin du processus que l'on déclarera l'index "complet" et "bon pour le service". Le problème, c'est que certaines des pages sont déjà obsolètes à la fin du processus. D'autres auront disparu. De nouvelles pages auront été mises en ligne, et ne figurent pas dans l'index.

Bref, ce n'est pas un moyen très efficace de gérer le problème de la fraîcheur des pages. Google, qui lançait un "full crawl" par mois, se retrouvait donc avec un index dont la fraîcheur posait problème. La solution adoptée pour compenser le problème consistait à lancer un robot différent "le fresh crawler" destiné à détecter de nouvelles pages apparues entre deux "full crawl".

De nombreux chercheurs ont donc réfléchi à des solutions alternatives ou à des perfectionnements de la méthode. La plupart des travaux ont abouti à la création d'agents d'indexation spécialisés, extrêmement efficaces pour "aspirer" et "surveiller" des portions limitées du web. Parmi ces travaux, on peut citer les techniques de crawling ciblé (focused crawling) et de crawl intelligent (intelligent crawl), ou les techniques de crawl s'appuyant sur des algorithmes génétiques.

Mais la taille de la Toile dans son entier pose des problèmes particuliers, et les pistes efficaces pour résoudre le problème à cette dimension sont moins nombreuses.

Une solution : indexer plus souvent les pages les plus importantes

C'est la piste proposée par Cho et Garcia−Molina [1][2]. Les deux chercheurs américains sont partis du principe que la Toile représente un volume global de données qui atteint (c'est une estimation indirecte, valable au moment de la publication de l'article) plusieurs téraoctets de données. Gérer une telle quantité d'information représente déjà un défi gigantesque sur le plan technique. Mais toute cette masse de données n'est pas forcément utile pour la plupart des utilisations classiques d'un moteur de recherche, il n'est donc pas absurde de "limiter" la taille de l'index, de manière arbitraire, à une portion de cet ensemble, sans dégrader gravement la pertinence des résultats.

Cho et Garcia Molina se sont donc attachés à améliorer la qualité d'un index de taille fixe et limitée. Cette approche convient notamment si l'on souhaite développer un moteur performant, mais avec des ressources limitées. Elle privilégie la fraîcheur et la qualité de l'index, par rapport à l'exhaustivité.

Le fondement de cette approche repose sur une évaluation de l'importance des pages. En effet, la qualité d'un index limité reposera sur la possibilité d'indexer toujours les pages "importantes", et jamais les pages "sans importance".

Le problème est de définir les bons critères pour évaluer l'importance d'une page...

La technique proposée par Garcia et Molina repose sur une combinaison des critères possibles suivants pour attribuer à chaque url une note d'importance :

1°) La similarité avec une requête donnée (évaluation sémantique)
2°) Le décompte des backlinks pointant vers cette page
3°) Le pagerank (dans ce cas, tous les backlinks n'ont pas la même valeur)
4°) Le décompte des forwardlinks
5°) La localisation des pages (note dépendant du TLD, de la position dans l'arborescence)

Une fois que l'on a défini une fonction d'évaluation de l'importance de la page, on peut s'en servir pour déterminer de manière plus intelligente l'ordre de crawl pour les pages, en indexant toujours de manière prioritaire les pages dotées de la meilleure note d'importance...

A noter que les travaux de Cho et Garcia−Molina ont montré que s'il ne fallait se baser que sur un seul de ces critères, le pagerank fournissait l'évaluation la plus efficace.

Le crawler incrémental

Par la suite, les deux chercheurs ont travaillé sur une application de cette fonction d'évaluation des pages à un crawler "incrémental". Un tel robot d'indexation, par opposition à la technique d'indexation par lot, ne s'arrête jamais. Il ne s'agit plus d'aspirer tout le web et de s'arrêter, pour recommencer à zéro plus tard. Un crawler incrémental a pour mission de déterminer quelles pages sont susceptibles d'être devenues obsolètes, et de les mettre à jour.

Cette approche incrémentale est beaucoup plus économique en ressources, car elle évite en théorie d'avoir à crawler des pages qui n'ont pas changé. Et il devient possible d'adapter la périodicité des crawls à la fréquence de changement des pages d'un site donné. Le cycle de crawl n'est plus uniforme : il varie d'un domaine à un autre.***

Une telle technique n'est utilisable que si l'on a pris soin d'évaluer préalablement la fréquence de changement des pages dans un site donné. Et il faut séparer ici deux évolutions différentes : la disparition de pages et l'apparition de nouvelles d'une part, et les changements de contenu dans une page donnée d'autre part. Un moteur de recherche qui décide de s'appuyer sur les techniques incrémentales a donc besoin, pour que l'indexation soit pertinente, d' attendre un certain délai pour collecter des informations fiables sur la manière dont l'index évolue. Il convient par la suite de mettre à jour ces informations, et de lancer, périodiquement, un "batch crawling" général pour éviter une dérive de l'index.

L'ordre de crawl, dans le robot d'indexation proposé par Garcia−Molina et Cho, est donc finalement déterminé en fonction de deux groupes de critères différents :

− l'importance des pages

− et leur probabilité d'obsolescence...

Un crawler incrémental, sur une base ouverte (L'exemple de Webfountain)

Jenny Edwards, une spécialiste australienne des robots d'indexation, a proposé une méthode similaire, mais capable de fonctionner non pas sur un index limité, mais sur un index ouvert, et de taille variable[3][4]. Ses travaux ont trouvé leur application dans le nouvel outil de recherche d'IBM : Webfountain, avec le crawler baptisé "Seeker".

Pourquoi un index ouvert ? Parce qu'au sein de "Big Blue", Jenny Edwards savait qu'elle pouvait disposer de stations de travail ultra−puissantes, d'une bande passante très importante, et de capacité de stockage sans équivalent. Travailler sur une solution flexible, capable de suivre la croissance de la Toile n'était plus une utopie avec de tels moyens.

Dans ce cas, l'adjectif "incrémental" prend une signification légèrement différente. Il ne s'agit plus de mettre à jour un index fixe, mais de mettre à jour les pages existantes de l'index.

Seeker ne tient pas compte de l'importance des pages, mais uniquement de leur rythme de changement. Sa particularité est de "séparer" le traitement des urls, en fonction de leur fréquence de mise à jour... Ce qui permet de traiter différemment les sites d'actualités, des sites qui ne disposent que de pages fixes rarement modifiées.

Vers le niveau de fraîcheur optimal ?

Toutes ces techniques évoluées permettent d'améliorer le niveau de "fraîcheur" des index.

Si l'on considère la Toile dans son ensemble, disposer d'un index parfaitement à jour reste toutefois encore une utopie. Mais pour celui qui veut développer un moteur d'actualités, la démonstration est faite qu'il est parfaitement possible d'élaborer un crawler capable de maintenir son index à jour, même si elles s'avèrent changer plusieurs fois par heure.

Les techniques de crawl évoluées dotent les moteurs de recherche, jadis un peu myopes, de capacités d'indexation désormais très impressionnantes... C'est une avancée importante dans la problématique de la pertinence des résultats. Mais c'est dans l'amélioration de tous les autres maillons de la chaîne de production de ces résultats que se jouera l'avenir des moteurs de recherche ...

** Est−il possible de crawler toutes les pages de la Toile ?*
Si on remonte à deux ou trois ans seulement, il était communément admis que la croissance du WWW était exponentielle, et que face à cette explosion, indexer l'ensemble de la Toile était totalement illusoire. Aujourd'hui, les données ont un peu changé.
La puissance des machines d'indexation continue de croître selon la loi de Moore et la bande passante utilisable a également progressé dans des proportions notables. Surtout, la capacité des supports de stockage magnétique s'est développée de manière extraordinaire, et l'on peut faire tenir plusieurs dizaines de Teraoctets dans un volume de la taille d'une commode !
Dans le même temps, la vitesse de croissance du nombre des pages webs n'est plus du tout

exponentielle ! Les dernières évaluations laissent penser que le rythme de cette croissance diminue.
Bref, il y a encore un avenir pour les moteurs dont l'ambition est d'indexer toutes les pages de la Toile accessibles par les techniques de crawl actuelles ou futures... Reste à savoir si un moteur dont l'index est de la taille du web est forcément utile ou plus pertinent. Rien n'est moins sûr, et le débat n'est pas près d'être clos...

*** Une solution classique à ce problème est celui du " round robin logiciel " : les urls à crawler ne sont pas choisies n'importe comment, un programme choisit un groupe de sites à crawler, et au lieu de visiter toujours le même site, change le site crawlé à chaque fois, pour espacer les visites.*

**** Le comportement d'un robot d'indexation incrémental décrit ici correspond trait pour trait à celui de Googlebot depuis quelques mois. Ce n'est probablement pas une coïncidence : le robot de Google semble fonctionner de manière incrémentale.*

Bibliographie :

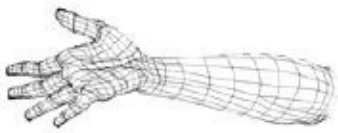
[1] The Evolution of the Web and Implications for an Incremental Crawler
Junghoo Cho, Hector Garcia−Molina / Department of Computer Science Stanford, CA 94305
December 2, 1999

[2] Parallel Crawlers
Junghoo Cho / University of California, Los Angeles − Hector Garcia−Molina / Stanford University
2002

[3] Webfountain d'IBM : un moteur de recherche révolutionnaire
Philippe YONNET
novembre 2003

[4] An Adaptive Model for Optimizing Performance of an Incremental Web Crawler
Jenny Edwards, Kevin McCurley, John Tomlin
25 février 2002

Conclusion



ATTENTION
La liste des abonnés n'est pas disponible. Nous ne la donnons ni ne la revendons à personne. En revanche, vous pouvez sponsoriser notre lettre d'information : contactez−nous pour discuter des modalités.

les crédits

Chronique sur « Google »	Olivier Duffez de **WebRankInfo**
Chronique sur les moteurs	Gilbert Wayenborgh de **DeepIndex**
Chronique sur les entreprises	Jean−Claude Benard de **Quadramultimédia**
Chronique sur la visibilité	Damien Guigue de **Yooda**
Chronique sur le référencement marketing	Gautier Girard de **Cerise Rouge**
Humour et Internet	Frédéric Lepage de **Les BD qui bougent**
Chronique sur le référencement	Marie Pourreyron d'**Altitude Référencement**
Chronique sur l'aspect commercial	Pierre Frigon de l'**Agence Hyperclics Marketing**
Infodesign	Max, le talentueux !
Le reste (mise en page, édito, actualités, les nouveautés et ...)	CHRis Hédé de **MoteurZine / Enfin**