Chronique

de la recherche sur le web

par Marc Duval 3 décembre 2001, vol.1, n° 10

Les secrets de Google

Google est un automate de recherche qui a ses petits secrets. En voici quelques-uns.

Google tronque énormément

La troncature est une fonction du langage de recherche qui permet de rechercher la déclinaison d'un mot avec un symbole placé devant, derrière ou à l'intérieur d'un mot. Le symbole peut être « * », « ? » ou « $ ». Par exemple, cherch* trouvera chercher, cherché, cherchez, chercheur... ; *cherche trouvera recherche, év*nement trouvera évènement et événement. Google n'a pas cette fonction. Elle doit être remplacée en utilisant l'opérateur booléen « OR » si l'on veut que l'un ou l'autre terme soit présent dans une page. Par exemple, si l'on veut éliminer la déclinaison d'un mot (singulier et pluriel) ou d'un verbe, il faut ajouter tous les mots précédés du signe moins «–». L'utilisation du formulaire de recherche détaillé permet de ne pas se préoccuper de ce détail. Google a cependant un talon d'achille : il n'accepte que 10 termes de recherche.

Les Google nationaux

Google a développé 67 pages d'accueil nationales. La plupart d'entre elles sont des traductions de la page d'accueil en anglais et 23 ont l'option de recherche dans la zone géographique nationale et internationale dans la langue de la page d'accueil. Par exemple, le Canada a deux pages d'accueil -une en français et l'autre en anglais avec une adresse : http://www.google.ca . 3 zones de recherche sont disponibles : le web, le web francophone et le web canadien.

Les pages cachées

Google a cette particularité de permettre la consultation de ses pages cachées. Cette fonctionnalité est active seulement sur son site et inactive sur d'autres comme sur celui de Yahoo. Or certains sites listés dans le résultat de recherche de Google ont une ligne de programme de redirection vers une page avertissant que la page recherchée est destinée aux membres.

S'il vous arrive de rencontrer une telle page, pointez votre curseur sur « copie cachée », cliquez sur « Activer un lien dans Composer » (pour Netscape et IE6). Vous pourrez alors lire cette page fugitive.

Les sites fantômes

Google recense des sites fantômes. Le plus bel exemple est celui du Journal de Montréal et du Journal de Québec. Habituellement, lorsque l'on fait une recherche sur un site en particulier (site:www.dsi-info.ca), Google sortira toutes les pages indexées du site. Or, les 2 journaux ont cessé de publier sur le web depuis l'an 2000. Ces pages ne sont plus dans l'index. Avec la requête « sont site:www.journaldemontreal.com » que trouve-t-on ? Des pages avec l'adresse « www.canoe.qc.ca » qui ne sont pas des

redirections.

Les documents cachés

Depuis quelques mois, Google a ajouté le champ de recherche «filetype». En utilisant le formulaire de recherche détaillé, vous pouvez chercher 6 types de fichiers : Adobe Portable Document Format (pdf), Adobe PostScript (ps),Microsoft Excel (xls), Microsoft PowerPoint (ppt), Microsoft Word (doc), et Rich Text Format (rtf). En plus de ces types de fichiers, il est possible de rechercher des fichiers Microsoft Works (wks, wps, wdb), Microsoft Write (wri), Text (ans, txt) Adobe Photoshop (psd), Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku), Lotus WordPro (lwp), MacWrite (mw), DBASE 3 (dbf), Paradox (db), WordPerfect... et davantage. De quoi amuser les curieux et surprendre les naïfs et les négligents qui mettent sur des serveurs publics des documents privés. Conseil d'ami : si votre fournisseur d'accès internet vous donne un espace disque pour votre page personnelle, ne l'utilisez pas pour faire la sauvegarde de votre disque dur. Même le fichier «robots.txt» ne vous préservera pas des regards de Google, dixit Google lui-même. Autre conseil : il est notoire que les virus s'attaquent davantage aux produits de Microsoft. Dans tous les cas, n'ouvrez pas les fichiers autres que ceux en HTML ou sauvegardez-les et vérifiez-les à l'antivirus.

Les pages tronquées

Vous pensez peut être qu'en cherchant sur Google, vous faites une recherche sur les pages au complet. Détrompez-vous. Google s'arrête de chercher dès qu'il atteint la taille de fichier de 101k. Comparez, ce document http://www.geocities.com/CapitolHill/Senate/5428/maitrise.html a 101 k dans Google et 388.7 dans AllTheWeb.

Google et la bourse

Vous pouvez rechercher des informations boursières avec Google en entrant le symbole boursier d'une entreprise enregistrée aux bourses de New York, Nasdaq et AME. La première référence affichera un graphique en vignette. En cliquant sur la vignette, vous êtes dirigé vers Yahoo Finance.

L'annuaire téléphonique de Google

Cherchez-vous une adresse aux États-Unis ? Tapez le prénom, le nom et la ville de votre Américain et vous aurez ses coordonnées et même la carte routière pour vous y rendre.

Les cartes routières de Google

De la même façon, si vous avez une adresse américaine, écrivez-la dans la case de recherche. Google affichera une carte en vignette avec un lien vers Yahoo! Maps et MapQuest

Google fête

Vous avez peut-être déjà vu une page d'accueil de Google lors d'une occasion spéciale. Google participe aux différents évènements en affichant un logo de circonstance. Surveillez ces dates : 31 octobre (Halloween) ; 4e jeudi de novembre (l'Action de Grâce) ; 1er janvier (Nouvel An) ; 2 février (Jour de la marmotte) ; 14 février (St-Valentin) ; 17 mars (La Saint-Patrick) ; 1er avril ; Pâques ; 14 mai (fête de mères) ; 18 juin (fête des pères) ; 1er juillet - 4 juillet - 14 juillet (fêtes nationales du Canada, des États-Unis et de la France)... Les logos sont affichés à cette adresse : http://www.google.com/holidaylogos.html.

Google se multiplie

Pas satisfait d'avoir une adresse URL, Google en a trois autres : http://www1.google.com, http://www2.google.com, http://www3.google.com. Ces adresses correspondent aux trois centres de données de Google : Santa Clara (Californie), Sunnyvale (Californie) et à Herndon (Washington, DC). Ce dernier centre est une assurance contre les tremblements de terre ou autres catastrophes qui pourraient

survenir sur la côte ouest des États-Unis. Ces trois centres de données sont identiques.

Google dans votre fureteur

Google offre de mettre sur votre fureteur une « barre d'outils ». Cette barre de recherche est disponible seulement sur Internet Explorer. Elle est aussi un excellent moyen pour Google d'intercepter vos recherches pour en établir des statistiques. Si vous ne souhaitez pas être espionné par Google, naturellement vous pouvez mettre un lien dans vos favoris. Toutefois, Google propose 3 boutons à insérer dans votre fureteur : Google Search (met en évidence des termes de recherche), Google Scout (pages similaires) et Google.com.

Google traduit

Vous avez sans doute rencontré des références suivies de la note « Traduire cette page ». Si vous connaissez l'adresse d'une page, vous pouvez aussi vous diriger vers sa page d'outils linguistiques qui comprend un module de traduction.

Les nouvelles sauce Google

Depuis octobre 2001, Google a intégré la mise à jour de sites de nouvelles sélectionnés dans un délai de 24 heures. Le hic est bien *l'intégration* dans sa banque de données sans distinction contrairement à ce que font Northern Light et AllTheWeb. Ces nouvelles ne peuvent bénéficier de la technique de classement PageRank pour les faire apparaître, ni d'un affichage distinct. Pour séparer le bon grain de l'ivraie, il faut ajouter le mois et l'année encadrés de guillemets anglais comme ceci "décembre 2001" ; la fonction de recherche par périodes de date ou par dates n'étant pas adaptée à la recherche de nouvelles.

Google a une équipe de chercheurs pour améliorer les capacités de recherche et le graphisme du site, et pour ajouter de nouvelles fonctionnalités. Les dernières trouvailles qui sont sur les planches sont l'affichage des pages cachées en vignettes, un classement selon la popularité d'un site d'après les internautes et une collection payante de documents comme celle de Northern Light. La banque de données de Google devient de plus en plus un capharnaüm où seuls les sites commandités sont mis en évidence. Il lui faut un nouveau classement par types de documents bien identifié. Il pourrait apparaître subdivisé ainsi : une partie répertoire, une partie nouvelles*, d'autres en pages web et pdf - et autres formats populaires. D'un site de recherche réputé simple à sa sortie, Google se dirige vers un portail de recherche dont la complexité sera accessible aux seuls experts.

*Google a ajouté quelques lignes d'actualités au début des résultats de recherche en anglais le 11 décembre 2001.

Marc Duval
Service de recherche documentaire DSI
1280-4, bd de Montarville.
Boucherville. Québec. Canada
J4B 8B4
marcduval@dsi-info.ca