Les outils de recherche d'images (1ère partie)

Retour au sommaire de la lettre

Nous débutons dans cette lettre une série d'articles sur les outils de recherche spécialisés sur certains formats spécifiques de fichiers ou de données : image, actualité, fichiers PDF, etc.

Nous inaugurons cette série avec les outils de recherche d'image. Ces sites seront étudiés en fonction des types de données qu'ils proposent. Ce mois-ci, ce sont les moteurs de recherche d'images proposés par les leaders de la recherche d'information généraliste (Google, Fast, Lycos...) qui sont comparés. Le mois prochain, nous étudierons plutôt les banques de données d'images et moteurs spécialisés (Corbis, Ditto, etc.). Enfin, un article sera plus spécifiquement consacré aux métamoteurs du domaine.

Comment les moteurs de recherche d'images effectuent-ils leurs investigations ?

Les moteurs de recherche identifient la plupart du temps les images qu'ils proposent à l'intérieur des pages détenues dans leur index web. Lorsque vous tapez le mot clé "napoléon", le moteur tente de retrouver les "bonnes" images selon plusieurs critères :

- Le nom du fichier : il proposera par exemple des images ayant pour nom napoleon.gif ou napoleon.jpg.

- Le descriptif textuel de l'image : les webmasters, lorsqu'ils créent leur site, ont la possibilité de placer, dans le code HTML de leurs pages, un petit texte décrivant chaque image affichée (option "ALT" de la balise <IMG>). Ce texte s'affiche, sur certains navigateurs, lorsqu'on passe la souris sans cliquer sur l'image ou avant le téléchargement de cette dernière. Ce texte descriptif est pris en compte par le moteur pour effectuer ses recherches. Exemple :
<IMG SRC="images/nap.jpg" ALT="Napoléon à Waterloo">
L'image ci-dessus sera donc potentiellement identifiable pour les mots clés "Napoléon" et "Waterloo".

- Le texte "autour" de l'image. Le moteur tient compte du texte de la page "proche" de l'image. Si celui-ci contient le mot "napoléon", l'image pourra être prise en compte même si son nom n'est pas évocateur (exemple : nap28.gif). Si un lien textuel permet d'afficher une image, le texte du lien sera également très important.

- Les balises Meta. AltaVista, par exemple, peut prendre en compte le texte des balises Meta pour retrouver une image contenue dans une page.

Comparaison de 6 outils de recherche d'images

Etudions donc de plus près la fonction "Images" des moteurs de recherche leaders du Web. Pour chacun d'entre eux, nous avons essayé d'indiquer des critères quantitatifs, qualitatifs et fonctionnels pour tenter de mieux les comparer. Nous avons testé dans cet article les moteurs "images" de :
- Google (http://images.google.fr/)
- Fast/AllTheWeb (http://www.alltheweb.com/?cat=img&cs=utf-8&q=&_sb_lang=fr+en)
- Lycos (http://www.recherche.lycos.fr/)
- AltaVista (http://fr.altavista.com/image/default)
- Tiscali recherche (http://www.nomade.tiscali.fr/)
- Et le nouvel outil de Yahoo! US (http://new.search.yahoo.com/images), en ligne depuis quelques jours.

Nous n'avons pas intégré les outils de recherche suivants :
- MSN France : une recherche d'images est possible dans la zone de recherche avancée (http://search.msn.fr/advanced.aspx), mais uniquement sur le nom des fichiers. Cette recherche semble, de toutes façons, ne pas fonctionner : les résultats sont identiques à ceux d'une recherche Web...
- Voila.fr : pas de test non plus, pour une raison identique. La recherche avancée (http://options.ke.voila.fr/plus_voila.php) ne propose qu'une option permettant d'effectuer des recherches sur des pages Web contenant des images, pas sur les images elle-mêmes.

- AOL France : le portail ne propose pas d'outil de recherche d'images.

Pour effectuer nos comparatifs, nous avons pris en compte 12 mots clés ou expressions : napoléon, cathédrale strasbourg, papillon, madonna, ferrari, bush, logo ibm, clé, tatouage, harry potter, lance armstrong et mona lisa.

Notons enfin que, si nous avons testé 6 outils de recherches différents, seules trois technologies sont utilisées par les moteurs : celles de Google (Google + Yahoo!), AltaVista (AltaVista) et Fast (AllTheWeb + Lycos + Tiscali Recherche).

Comparatif quantitatif

Dans un premier temps, nous avons tapé ces mots clés sur chaque outil de recherche et avons noté le nombre de résultats proposé par chaque outil :

Moteur :	Google		altavista	fast	LYCOS	TISCALI RECHERCHE
Technologie fournie par :	Google	Google	altavista	fast	fast	fast
napoléon	11 400	11 400	2 977	565	537	**12 370**
cathédrale strasbourg	**323**	**323**	156	48	16	42
papillon	**15 400**	**15 400**	10 504	6 603	1 698	6 603
madonna	57 400	57 400	**63 400**	31 300	10 710	31 300
ferrari	**108 000**	**108 000**	91 098	57 376	13 580	57 376
bush	178 000	178 000	**371 870**	75 203	21 201	75 203
logo ibm	**6 690**	**6 690**	241	862	111	0
clé	**18 800**	**18 800**	2 775	1 221	202	16 453
tatouage	**1 940**	**1 940**	1 112	727	72	727
harry potter	**87 100**	**87 100**	35 326	20 979	5 487	20 979
lance armstrong	4 550	4 550	**4 718**	1 632	417	1 632
mona lisa	**5 220**	**5 220**	3 966	3 106	931	3 106
Moyenne :	**18,78**	**18,78**	11,95	6	1,58	8,70

Remarque : pour son moteur de recherche d'images, AltaVista positionne un OU par défaut comme opérateur booléen. Ainsi, la requête "cathédrale strasbourg" (sans les guillemets) est équivalente à "cathédrale OU strasbourg". Les autres moteurs positionnent un ET par défaut à la place d'un espace. Pour comparer des requêtes comparables, nous avons donc saisi sur AltaVista, sur ce test uniquement, les requêtes à deux termes avec un signe "+" : +cathédrale +strasbourg.

Pour donner une note à chaque moteur, nous avons attribué, pour une requête donnée, la note de 20 au moteur donnant le plus grand nombre de résultats. Puis, une règle de trois nous a permis de calculer les notes des autres moteurs.

Premier constat : Google fournit à Yahoo! ses résultats "tel quel", puisque les mêmes chiffres sont trouvés dans les deux cas (mais pas dans le même ordre, comme nous le verrons par la suite). En revanche, AllTheWeb, Lycos et Tiscali Recherche ne semblent pas travailler sur le même index, ou en tout cas avec les mêmes critères de tri, car les nombre affichés sont tous différents pour une même requête. Cependant, Tiscali Recherche semble assez proche des résultats d'AllTheWeb, alors que Lycos France s'en éloigne assez fortement en règle générale.

Globalement, c'est Google qui semble proposer le plus gros index de fichiers images sur le Web à l'heure actuelle, même si les chiffres affichés peuvent difficilement être vérifiés.

Comparatif qualitatif

Pour comparer de façon qualitative les 6 outils, nous avons tapé les mêmes 12 requêtes sur chacun d'eux. Puis, nous avons évalué les 20 premières réponses (les 20 premières images affichées) et noté combien, parmi elles, étaient pertinentes. La notion de pertinence pouvait être assez générale, en fonction du mot clé. Par exemple, nous avons considéré comme pertinentes, pour le mot clé "bush", des photos de Georges Bush père et fils, de la chanteuse Kate Bush, ou de toute personne portant ce nom. En revanche, pour le mot clé "papillon", nous n'avons pas accepté les photos du chien portant ce nom, car nous estimons que le moteur doit être capable de remettre la requête dans son contexte et nous proposer une photo extraite d'une page parlant, de façon globale, des papillons, thème certainement plus "populaire" que les chiens s'appelant ainsi. Nous avons parfois tenu compte d'un parti pris, comme pour Madonna, mot clé pour lequel nous n'avons noté que les images représentant la chanteuse. Mais ce même parti-pris a été pris en compte sur tous les outils.

Moteur :	Google		altavista	fast	LYCOS	TISCALI RECHERCHE NOMADE.FR
Technologie fournie par :	Google	Google	altavista	fast	fast	fast
napoléon	17	14	**19**	15	17	9
cathédrale strasbourg	**16**	**16**	12	10	**16**	12
papillon	7	6	**8**	4	1	4
madonna	7	8	4	7	**15**	7
ferrari	19	19	**20**	**20**	**20**	**20**
bush	18	17	**19**	7	9	8
logo ibm	17	10	17	2	**20**	0
clé	0	1	2	1	**3**	1
tatouage	13	16	**19**	16	13	16
harry potter	17	18	18	18	**20**	16
lance armstrong	**20**	19	19	**20**	18	**20**
mona lisa	**18**	**18**	16	2	17	2
Moyenne :	14,08	13,50	**14,42**	10,17	14,08	9,58

Dans ce domaine, c'est AltaVista qui obtient la meilleure note. Ses notes moyennes sont plutôt bonnes, voire excellentes. Les résultats de Lycos, s'ils ne sont pas nombreux (voir comparatif précédent), sont cependant plutôt pertinents puisque ce moteur se classe à la deuxième place, ex-aequo avec Google.

Notons que Yahoo!, s'il propose le même nombre d'images que Google, ne les affiche pas dans le même ordre. Il semble, du coup, légèrement moins pertinent en règle générale... Mais l'outil n'a que quelques jours d'ancienneté. Google doit certainement proposer des outils d'affinage et de tri de ses propres résultats, et Yahoo! doit peut-être encore avoir à affiner ses filtres...

Les résultats de Tiscali Recherche sont, là encore, très proches de ceux d'AllTheWeb.

Comparatif fonctionnel

Moteur :	Google		altavista	fast	LYCOS	TISCALI RECHERCHE
Technologie fournie par :	Google	Google	altavista	fast	fast	fast
Syntaxe avancée	OR, -, guillemets, fileype:	OR, -, guillemets, fileype:	+, -, guillemets	+, -, guillemets, fileype:, site:	+, -, guillemets	+, -, guillemets, fileype:, site:
Recherche avancée / filtres	Taille, format, couleur, domaine, filtre parental (US)	Taille, format, couleur, domaine, filtre parental (US)	Type, couleur, source, filtre parental	Type, couleur, transparence, filtre parental	Non	Non
Informations données	nom, taille, poids, url	nom, taille, poids, url	nom, taille, poids, lien	nom, taille, poids, lien	nom, taille, poids, lien, lien vers aperçu	nom, taille, poids, lien vers pop-up "infos"
Moyenne :	18	18	14	**20**	10	12

Nous avons noté ici, les facilités de recherche (syntaxe spécifique, présence d'une zone de recherche avancée, de filtres, etc.) ainsi que les informations fournies pour chaque image. Pour ce comparatif, c'est Fast qui sort gagnant avec AllTheWeb, moteur sur lequel on peut effectuer de nombreuses requêtes très précises (format, couleur / monochrome, recherche sur un site donné, etc.). En revanche, Lycos France et Tiscali Recherche n'offre pas de réelle zone de recherche avancée pour les images. Google est assez proche. En revanche, on espérait mieux d'AltaVista, la syntaxe avancée qu'il propose sur le Web (host:, etc.) ne semblant pas fonctionner sur le moteur d'images.

Récapitulatif

Moteur :	Google		altavista	fast	LYCOS	TISCALI RECHERCHE
Technologie fournie par :	Google	Google	altavista	fast	fast	fast
Quantitatif	18,78	18,78	11,95	6	1,58	8,70
Qualitatif	14,08	13,50	14,42	10,17	14,08	9,58
Fonctionnel	18	18	14	20	10	12
Moyenne :	**16,24**	15,95	13,7	11,59	9,94	9,97
Les "Plus"	Le nombre de résultats trouvés, la pertinence.		La bonne pertinence des résultats.	La syntaxe d'interrogation et la recherche avancée très complètes	Les résultats différents de ceux d'AllTheWeb, permettant une alternative de recherche.	Le "pop-up" contenant des informations sur l'image.
Les "Moins"		Légèrement moins pertinent que	La syntaxe avancée, assez faible	Le faible nombre de résultats	Le faible nombre moyen de	L'absence de recherche avancée

		Google. Pourquoi utiliser Yahoo! si les résultats de Google en direct sont meilleurs ?	pour un moteur dont c'est pourtant le point fort sur les recherches Web.	trouvés.	résultats, l'absence de recherche avancée.	

Pour calculer la note finale, nous avons donné un coefficient 1 à la note des comparatif quantitatif et fonctionnel et un coefficient 2 à la note du comparatif qualitatif, car celui-ci nous semble le plus important.

Google sort donc vainqueur de ce comparatif général, avec son "client" Yahoo!. Puis vient AltaVista, qui propose l'un des plus anciens moteurs d'images sur le Web. Fast semble un peu en retrait pour ce qui est de la recherche d'images. En revanche, il possède la meilleure syntaxe d'interrogation. Avec un index un peu plus "musclé", il devrait donc pouvoir se positionner en challenger de Google assez rapidement...