Les outils de recherche de fichiers PDF

Retour au sommaire de la lettre

Nous continuons notre série d'articles sur les outils de recherche spécialisés sur certains formats spécifiques de fichiers ou de données : image, actualité, fichiers PDF, etc. Au cours des mois précédents, nous avons étudié les moteurs de recherche d'images avec les outils de recherche traditionnels (Google, Yahoo!, Lycos, etc.), les outils spécialisés (Ditto, PicSearch, Corbis) et les métamoteurs (Ithaki, Mamma, Fazzle, Ixquick et Metahoo).

Ce mois-ci, nous étudierons les outils de recherche qui permettent d'effectuer des requêtes spécifiquement sur les fichiers PDF.

Nous avons sélectionné les moteurs de recherche suivants :

- **Google** (http://www.google.fr/). Sa recherche avancée (http://www.google.fr/advanced_search?hl=fr) propose le filtre intitulé "Limiter les résultats au format de fichier", puis le choix "Adobe Acrobat PDF (.pdf)". Il est également possible d'utiliser la fonction **filetype:pdf.** Exemple : confidentiel filetype:pdf

- **Fast/AllTheWeb** (http://www.alltheweb.com/). Sa recherche avancée (http://www.alltheweb.com/advanced) propose le filtre intitulé "File Format - Only find results that are", puis le choix "Adobe PDF (.pdf)". Comme pour Google, il est également possible d'utiliser la fonction **filetype:pdf**. Exemple : confidentiel filetype:pdf

- **AltaVista** (http://www.altavista.fr/). Sa recherche avancée (http://fr.altavista.com/web/adv) propose le filtre intitulé "Type de fichier :", puis le choix "Seulement fichier PDF". Comme pour Google et AllTheWeb, il est également possible d'utiliser la fonction **filetype:pdf**. Exemple : confidentiel filetype:pdf

- **Voila** (http://www.voila.fr/). Pas de possibilité avec la recherche avancée, mais il est possible d'utiliser la fonction **path:pdf** pour restreindre sa recherche à ce type de fichier. Exemple : confidentiel path:pdf

- **SearchPDF** (http://searchpdf.adobe.com/). Moteur de recherche "officiel" d'Adobe.

La recherche avancée d'Inktomi, via HotBot (http://www.hotbot.com/adv.asp?prov=Inktomi&tab=web) propose le filtre intitulé "Page content", puis le choix "PDF (Acrobat)". Mais les résultats retournés sont en fait les pages qui contiennent un lien vers un document PDF. Le filtre proposé n'est donc pas spécifiquement sur le format PDF lui-même, c'est pourquoi nous n'avons pas inclus ce moteur dans notre comparatif.

Récapitulatif des fonctionnalités de recherche

	Google	fast	altavista	voila	Adobe
Recherche avancée	http://www.google.fr/advanced_search?hl=fr Filtre "Limiter les résultats au format de fichier", puis choix "Adobe Acrobat PDF (.pdf)".	http://www.alltheweb.com/advanced Filtre "File Format - Only find results that are", puis choix "Adobe PDF (.pdf)".	http://fr.altavista.com/web/adv Filtre "Type de fichier :", puis choix "Seulement fichier PDF".	Pas de possibilité	Mots clés à saisir directement sur la page d'accueil (moteur spécifique du format PDF)
Syntaxe spécifique	filetype:pdf	filetype:pdf	filetype:pdf	path:pdf	non

Comparatif quantitatif

Dans un premier temps, nous avons tapé 12 mots clés, en français et en anglais sur chaque moteur de recherche et avons noté le nombre de résultats proposé par chaque outil. Pour chacune de ces interrogations, nous avons pris en compte les options les plus larges (toutes les langues, index mondial par exemple) :

Mots clés	Google	fast	altavista	voila	Adobe
confidentiel	21 300	2 115	10 372	1 625	42
étude	402 000	84 458	204 750	21 300	0
catalogue	188 000	72 936	203 549	5 530	3 309
guide	2 880 000	615 773	1 962 401	18 982	57 824
documentation	2 280 000	337 356	1 095 328	16 518	23 187
horaires	121 000	16 282	41 158	6 828	210
confidential	1 210 000	145 588	407 884	102	4 763
study	3 360 000	690 043	2 936 492	2 277	77 859
catalog	1 520 000	130 638	423 039	237	16 960
document	3 450 000	641 672	2 444 847	33 725	90 820
presentation	2 190 000	384 013	1 552 642	31 119	18 995
copyright	1 900 000	355 638	1 482 946	3 434	26 309
Moyenne :	19,87	3,54	11,98	0,42	0,23

Première constation : le moteur "officiel" d'Adobe semble faible, quantitativement parlant, et notamment en ce qui concerne les fichiers en langue française. Google semble imbattable à ce niveau, même si AltaVista, qui a augmenté son index dernièrement, fait bonne figure. AllTheWeb, de son côté est assez décevant avec des scores qu'on aurait imaginé bien meilleurs...
Le "cas" de Voila est assez spécifique puisqu'il se restreint aux fichiers en langue française, le moteur ne proposant pas, pour l'instant, d'index anglophone. Difficile, donc, de le comparer à des outils comme Google et AltaVista qui ont, eux, une vocation mondiale.

Comparatif qualitatif

Pour comparer de façon qualitative les 4 outils étudiés, nous avons tapé 12 mots clés et expressions (plus précis que ceux saisis dans le premier comparatif). Puis, nous avons évalué les 20 premières réponses (les 20 premier fichiers proposés) et noté combien, parmi elles, étaient pertinentes et en rapport avec les thèmes demandés.

Moteur :	Google	fast	altavista	voila	Adobe
statistiques internautes	16	13	12	3	9
étude de marché	17	5	18	11	3
catalogue jouets	17	0	12	1	0

nokia guide	20	16	20	4	19
documentation amortisseur	5	6	15	3	0
horaires avion paris	3	2	5	3	0
market study	19	19	20	0	15
microsoft word	20	2	0	7	0
salons internet	6	1	7	8	1
google pagerank	15	15	20	2	1
sars health	20	20	17	2	18
faq insurance	14	10	14	0	3
Moyenne :	17,21	10,55	16,97	5,68	6,20

Google et AltaVista sont assez proches en termes de pertinence, loins devant leurs challengers. On peut noter que sur la requête "piège" **microsoft word**, seul Google a proposé de nombreux fichiers traitant du logiciel-phare de Microsoft, tous les autres moteurs ont indiqué dans leurs résultats des fichiers traitant de thèmes divers, réalisés sous le logiciel Word avant d'être "traduits" en PDF. Ce simple exemple peut expliquer pourquoi Google est leader en termes de pertinence aujourd'hui sur le marché des moteurs de recherche...

Comme pour le classement précédent, la place de Voila est difficilement analysable, puisque le moteur est, de façon logique, assez peu pertinent sur des requêtes en langue anglaise.

Comparatif fonctionnel

Nous avons noté ici, les facilités de recherche (syntaxe spécifique) ainsi que les informations fournies pour chaque fichier.

Moteur :	Google	fast	altavista	voila	Adobe
Syntaxe avancée	filetype:	filetype:	filetype:	path:	Pas de syntaxe avancée
Informations spécifiques fournies	Version HTML, Pages similaires	Poids du fichier	Lien pour téléchargement "reader" PDF	Poids du fichier	L'url n'est pas indiquée
Note :	20	18	18	18	12

La plupart des moteurs proposent les mêmes informations, même si on aurait apprécié que Google affiche le poids du fichier, très intéressant dans le cadre de l'appréciation d'un futur téléchargement. En revanche, son utilitaire de conversion HTML "en ligne" est vraiment très intéressant et fait gagner beaucoup de temps. En revanche, l'interface utilisateur du moteur d'Adobe est très insuffisante : seul le titre (pas toujours expressif pour un document PDF) et un court résumé, pas toujours très lisible, sont proposés. Il faut cliquer sur le lien affiché (interne au site d'Adobe) pour obtenir plus d'infos sur le document en question, d'où une perte de temps non négligeable. A revoir...

Récapitulatif

Moteur :	Google	fast	altavista	voila	Adobe
Quantitatif	19,87	3,54	11,98	0,42	0,23
Qualitatif	17,21	10,55	16,97	5,68	6,20
Fonctionnel	20	18	18	18	12
Moyenne :	18,57	10,66	15,98	7,45	6,16
Les "Plus"	Important nombre de résultats, bonne pertinence. Traducteur HTML de fichiers PDF.		Bonne pertinence		
Les "Moins"	Manque le poids du fichier dans les résultats.	Assez peu pertinent. Faible nombre de résultats.	Moins de résultats que Google.	Restriction à la langue française	Moteur décevant sur tous les classements

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.

Le résultat donne, comme pour les outils de recherche d'images les mois précédents, **Google gagnant haut la main**. Il remporte les meilleures notes dans les trois classements, ce qui se passe de commentaires. Mais qui arrêtera Google ?

La place de Voila est à relativiser puisque ce moteur n'effectue ses recherches que dans l'univers du Web francophone. Peut-être sera-t-il intéressant de refaire ce test lorsque Voila aura lancé, puisqu'il semble que ce soit un projet en cours de réflexion chez France Telecom, un index anglophone digne de ce nom.

Enfin, si AltaVista est un digne challenger de Google, on reste un peu déçu par les performances d'AllTheWeb et du moteur d'Adobe, tous les deux assez peu performants et, en tout cas, à ne pas conseiller pour obtenir les meilleurs résultats possibles pour ce type de requêtes. Peut-être faudra-t-il à nouveau tester AllTheWeb lorsque sa fusion avec AltaVista sera effective, d'ici à la fin de l'année...