Chronique

de la recherche sur le web

par Marc Duval 5 mars 2003, vol. 3, n° 3

Une question de taille

Les automates de recherche comme Google ratisse le Réseau à la recherche de pages généralement web. Ils incorporent aussi des pages que des entreprises font inclure moyennant paiement. Ils acceptent aussi des soumissions gratuites de pages provenant de webmestres gérant de petits sites. En moins de dix ans, leurs banques de données sont passées de milliers de pages à des milliards tout en conservant et même en diminuant leur temps de réponses. Pour améliorer cette performance, on a dû faire une concession de taille: celle du fichier.

D'après une étude de la société Cyveillance en 2000[1], la taille moyenne d'une page web est de 10K. Une autre étude de la société @position[2] de 2003 note que la taille moyenne d'une page web est de 19K. C'est très peu pour donner un contenu substantiel. C'est pourquoi la majorité d'entre eux indexent les pages web jusqu'à 130K à l'exception d'Exalead qui arrête à environ 23K.

Google et Gigablast indexent jusqu'à 101k, Teoma à 115K, WiseNut à 120K, AltaVista et AntiSearch à 130K, Inktomi/HotBot/MNS Search jusqu'à 524K, OpenFind à 610K, AlltheWeb à 976K, Northern Light à 1418K et le Web francophone de Voila à 1582K.

La troncation des fichiers apporte deux conséquences importantes : la multiplication des pages et le silence.

La multiplication des pages

La troncation des fichiers par les automates de recherche influence directement la taille d'un fichier dès sa conception. Elle est d'autant importante lorsqu'un automate comme Google est prépondérant. La prépondérance de Google donne à la taille indexée une forme de norme. Pour qui connaît cette particularité, diviser un document en fonction de la taille est tentant. Cette division implique une multiplication des pages.

Le silence

La troncation des fichiers que font les automates peut être la cause de ce que l'on appelle le silence. Le silence est l'absence dans le résultat d'une recherche d'un document présent dans une banque de données. Il se produit lorsqu'un mot d'une requête se trouve au-delà de la taille de fichiers acceptés, par exemple des 101K chez Google. Les longs documents tels que des lois et des règlements, des thèses doctorales, des études, des articles encyclopédiques, des biographies sont sujets au silence.

En matière de recherche, la taille d'un fichier est un critère de sélection. Plusieurs automates l'ont mis dans leur format d'affichage soit par défaut, soit dans les préférences. En consultant la liste des références d'un résultat de recherche, l'internaute averti peut juger une page en raison de la taille du fichier. Il peut rejeter un fichier de moins de 5K ou rechercher des pages qui ont plus de 15K.

Seul AlltheWeb a l'option de la recherche par

---

[1] Brian H. Murray et Alvin Moore.(Consulté le 5 mars 2003). Sizing in the Internet: a white paper . [En ligne] Cyveillance, Arlington, VA, 2000. Page 4. http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf

[2] eStat/@position. (Page consultée le 5 mars 2003). «Google : Les chiffres en détail - mai/juin 2002» [En ligne]. La revue du référencement, février 2003. http://www.revue-referencement.com/etudes/etude_google.htm

la taille de documents. Cette option se trouve dans son formulaire de recherche. L'internaute peut chercher un fichier ayant une taille de valeur égale ou plus petite ou plus grande que celle choisie. Si l'internaute est expérimenté, il peut travailler avec le préfixe **filesize :** accompagné des signes = (égal), > (plus grand), < (plus petit) et [n;n] (intervalle de nombres).

Conclusion

La taille d'une page web est un critère primordial pour le temps d'affichage. Il peut être une raison pour laquelle un internaute coupe la communication et ne retourne pas sur cette page. A côté de cette considération « ergonomique», la taille d'un fichier est un facteur déterminant pour qu'une page soit trouvée à l'aide d'un automate de recherche. Si l'un des mots d'une requête est au-delà des 101K de Google, l'automate le plus populaire, la page demeurera inconnue. Seul AlltheWeb a une option de recherche pour la taille d'une page web d'autant qu'il indexe les fichiers jusqu'à 1Mo. Ce qui lui confère un avantage sur les autres sites de recherche. La taille d'un fichier détermine enfin le temps de réponse. La compression de données sera un enjeu important pour augmenter la taille d'indexation d'un fichier sans compromettre le temps de réponse dans la perspective de banques de données de plus de 4 milliards de pages.

Marc Duval
Bibliothécaire professionnel
Service de recherche documentaire DSI
1280-4, bd de Montarville.
Longueuil. Québec. Canada
J4B 8B4
marcduval@dsi-info.ca
http://www.dsi-info.ca