Objet :	**Recherche « avancée » sous Google**	Date : Version :	**02/02/2009** **1.0**

Google, le moteur de recherche le plus utilisé sur Internet, le plus consulté, probablement celui qui indexe le plus grand nombre de pages ou de données présentes sur Internet, le plus… incontournable.
Pour ce titre, Google est devenu le moteur de recherche de tout et n'importe quoi : au delà des recherches habituelles, on le retrouve utilisé aussi bien pour chercher des fichiers vidéo, audio licites (ou pas), comme pour rechercher des pages Web pouvant présenter des informations masquées, voire des vulnérabilités de sites web, ce qui amène progressivement sur ce qui est couramment nommé le Google Hacking.
Dans la continuité, Michael Sutton (ex-SpyDinamics) a estimé qu'à travers de simples requêtes Google, on détectait environ 21 % de sites sensibles au Cross Site Scripting, 14% pour le SQL Injection, 9.5 % pour le PHP include et 8% pour les Buffer overflows connus.

Comment çà marche ?

Comme n'importe quel moteur : un automate démarre par une URL qui lui est soumise (automatiquement ou non) puis explore l'ensemble des pages ou liens depuis cette URL, pour le même domaine. Cet automate est nommé un Crawler ou Robot, dans le cas de Google, il se nomme GoogleBot.
PageRank : valeur attribuée par Google en fonction du nombre de pages Internet qui pointent vers une site X . Plus le PageRank d'un site est élevé, plus ce site apparaîtra dans les premiers résultats d'une requête Google.

Recherches de base :

- L'ordre des mots clés est pris en compte ; si vous tapez `google hacking active directory` au lieu de `hacking active directory google`, les résultats seront différents parce que les 2 mots sont recherchés séparément.

- La recherche d'un bloc de mots ou d'une phrase se fait en encadrant le tout par des guillemets .
- Il n'y pas de distinction entre minuscules et majuscules

Les opérateurs :

Trois opérateurs booléens permettent de construire une requête sur mots-clés.

- ➔ AND exprimé par + (plus). Exemple : toto + titi renvoi les pages contenant ces 2 mots.
- ➔ OR exprimé par | (pipe). Exemple : toto | titi renvoi les pages contenant l'un des 2 mots.
- ➔ NOT exprimé par – (moins). Exemple : toto -titi renvoi les pages contenant toto mais surtout pas titi.

Tous ces opérateurs peuvent être combinés à souhait, dont les groupes sont à spéparer par des parenthèses.
Par exemple : `google+hacking –(« active directory »|annuaire)` cherchera les pages contenant `google` et `hacking` mais ne contenant pas la phrase `active directory` ni le mot `annuaire`.

Ajoutons d'autres opérateurs, non booléens et très puissants :

- ➔ `~ texte` demande les mots clés s'approchant ou synonymes (même français) de 'texte'. Exemple : ~info fournira les pages contenant un synomye tel que information, info, help, guide,..
- ➔ `chiffre1..chiffre2` renvoi les pages contenant un chiffre entre chiffre 1 et chiffre2. Exemple : Windows 1995..1997 renvoi des pages faisant référence à Windows 95.
- ➔ `mot1 * mot2` renvoi les pages contenant mot1 suivi de mots xx et dans la séquence finie par mot2. Exemple : « le petit » * rouge renverra des pages portants sur le comte de Perreau ou un livre Chinois ou….

Les filtres

Au delà des opérateurs booléens, il peut être utile (et surtout performant) d'utiliser des filtres pour affiner la requête. Chaque filtre peut-être associé à l'un des opérateurs ci-dessus. En voici quelques uns :

- ➔ **`site:`**`domaine.com` permet une recherche sur un domaine spécifié.
- ➔ `intitle:'titre de page web'` spécifie de ne prendre que les pages Web affichant le titre contenu dans <title> titre de page web</title>. Lorsque la page affichée contient un listage de répertoire/fichiers (genre FTP), le titre de la page est 'Index Of'.

- ➔ `ext:fic` recherche les fichiers portant uniquement l'extension fic (s'applique bien sûr à pdf, doc, htaccess,…..)
- ➔ `inurl :.ext` demande à Google de ne renvoyer que les pages contenant des liens vers des fichiers portants l'extension .et (s'applique à aussi .php, . html, .asp, .html,…)
- ➔ `intext :motclé` recherche dans les pages le mot clé passé en paramètre
- ➔

Exemples

1/ imaginons vouloir rechercher des sites web configurés volontairement (ou surtout pas) pour renvoyer la liste des fichiers et non les pages web qui s'y trouvent (traditionnellement index.html, index.php, default.htm,…)
Il suffirait de saisir : `-inurl(.php|.html|.asp|.htm) intitle :'Index Of' +'last modified' +'parent directory'`

2/ imaginons vouloir rechercher parmi les résultats ci-dessus, une liste de fichiers MP3
`-inurl(.php|.html|.asp|.htm) intitle :'Index Of' +'last modified' +'parent directory' +'inurl:.mp3`

Passage de paramètres via GET

Déviances

-Google Bombing : technique qui consiste à faire augmenter le PageRank d'un site pour attirer artificiellement (ou par tromperie) l'utilisateur. Par tromperie ? Avec des mots clés très prisés positionnés sur les pages externes qui pointent vers le site X, sachant que ces mots sont sans rapport avec le réel contenu du site X.

Outils et liens complémentaires

GoogleHack

Outils gratuit (hébergé chez Google Code) qui vise à simplifier les recherches particulières via une interface graphique

http://www.commentcamarche.net/telecharger/telecharger-34056788-google-hacks

Article HSC / Hakin9

http://www.hsc.fr/ressources/articles/hakin9_googlehacking/CA-Hakin9-06-2008-googlehacking.pdf