Pokročilé (vy)užívání googlu část 1

Vložil/a czokl [1], 12 Březen, 2006 - 14:28

- Programming [2]
- Security [3]

Množství informací, které se na internetu vysktují, nejsme schopni už dlouhou dobu pojmout. I proto dnes máme vyhledávače, které tyto informace třídí. A některé z nich toho při trošce naší snahy dokáží ještě mnohem více. Pokud Vás zajímá co přesně, čtěte dále.

-[intro]-

Google.com je asi nejkvalitnější fulltextový internetový vyhledávač. V době psaní článku má zaindexováno něco přes 8 miliard webových stránek, což je opravdu obrovské číslo. V tomto směru sice podle mých informací vede nyní yahoo(asi 20 miliard stánek), ale konec konců kvalita nad kvantitou povětšinou vítězí a zeptejte se kohokoli znalého v IT na nejlepší vyhledávač a odpověď bude myslím jasná. Ale v tomto prohlížeči se dají díky jeho pokročilým funkcím vlastnostem vyhledávat věci jako jsou přístupové informace do různých systémů, nelegálně distribuované multimediální soubory nebo třeba čísla kreditních karet. Takovému vyhledávání se říká "google hacking" a s jeho možnostmi, technikami a obranou proti němu bych Vás chtěl seznámit. Poznámka: konkurenční vyhledávače mají většinou stejnou nebo velmi podobnou syntaxi, takže můžete zkusit štěstí i jinde než přímo na google.

-[Upozornění]-

Abych předešel rejpalům hned v úvodu, možná se po přečtení následujících řádek nedozvíte nic nového, dle mého osobního názoru jsou metody gogole hackingu publikované dostatečně, ale zatím jsem nenašel na českém internetu kompletní průvodce touto problematikou a tak jsem se pokusil něco takového sám sepsat. Snad se aspoň trošku povedlo. A taky je dosti možné, že v době čtení článku bodu některé metody nefunkční, internet a s ním i google se neustále vyvíjejí, takže myslet si, že všechny dotazy budou fungovat v nezměněné podobě až do smrti by bylo dosti naivní. Proto jsem se také snažil hlavně vysvětlovat základní principy, po jejichž pochopení a následné chvilce testování bude sestavení dotazu už banalitou.

-[Obsah]-

-[Pokročilé vyhledávání]-

Každý z Vás na googlu určitě už něco někdy hledal a možná vám to ani nepřišlo příliš super, třeba když já jsem poprvé na google něco hledal, přišlo mi to stejné jako všude jinde. Rozdíl byl pouze v

množství vyhledaných stránek. Hlavní předností google je totiž pokročilé vyhledávání, kdy google dokáže podle určité syntaxe třídit výsledek. A to velice jemně. Jak na to? Můžete to udělat dvěma způsoby (dělení problému na dvě části si v tomto dokumentu ještě užijete), buď zadáte do prohlížeče http://www.google.com/advanced_search [4] (samozřejmě se tam dá prokliknout i z hlavní stránky google) a nebo budete syntaxi vypisovat přímo. Brzy přijdete na to, že druhý způsob je mnohem rychlejší a lepší, ale i první způsob se dá bez problému použít. Já se jím ovšem zabývat nebudu, protože je to stejně jakási grafická/klikací obdoba vypisování query (=dotaz, budu tak nazývat všechny dotazy/hesla pro vyhledávač) přímo do inputu (taková ta kolonka pro vstup). Myslím, že nyní je čas na představení klíčových slov.

Obecná syntaxe heslo:argument

inurl, allinurl

Argument tohoto hesla se musí objevit v url, například "inurl:admin/login.php" zobrazí všechny výsledky, kde se v adrese objeví toto spojení. pokud allinurl tak můžete dokonce vyhledávat více slov v jedné url, protože všechno za tímto slovem se musí v url vyskytnout. tedy například "allinurl: admin backup", ale pozor na pořadí slov, přehozením těchto dvou slov získáte jiný výsledek.

intitle, allintitle

Podobné, akorát že výsledky nehledá v url ale v titulku stránky. Jinak platí úplně ty samá pravidla. Např. "intitle:mp3" nebo "allintitle: free mp3 download".

intext, allintext

Zde je rozdíl pouze v tom, že vyhledává v textu. Pravidla opět stejná. Příklad "intext:password"

site

Aneb vyhledávání na určité doméně a můžete vyhledávat na doméně kteréhokoli rádu. například "site:org", "site:localstudio.org", "site:net-way.localstudio.org".

link

Hledáte všechny stránky, které mají na Vám zadanou stránku odkaz. Takže třeba "link:http://security-portal.cz" nalezne všechny stránky odkazující na s-p.cz. No zrovna v tomto případě google všechny stránky zaindexovány nemá, naše referers jsou o dost bohatší, ale co už :)

related

Aneb příbuzná stránka ke stránce, kterou zadáte. Ve slova smyslu obsahu samozřejmě. Zkuste třeba "related:www.securityfocus.com" a dostanete několik desítek stránek zabývajících se počítačovou bezpečností.

cache

Cached stránky jsou ty, které má google umístěny přímo na disku a jsou to vlastně "odlehčené" kopie stránek, které má indexovány. Toto je velice užitečné, nejednou mi to výrazně pomohlo. Zde je malinko změna, protože musíte zadat stránku. Například "cache:security-portal.cz novinky" zobrazí cache stránky s-p.cz a zvýrazní na ní slovo novinky.

filetype

Typ souboru, respektive koncovka souboru. Nemusíte hledat pouze htm a html stránky, ale i jiné datové soubory a o tom vlastně většina google hackingu je. Například "filetype:pdf"

""
Do úzovek se uzavírají přesné výrazy, tedy slovní spojení, které se musí vyskytovat na stránce

přesně tak, jak jsme zadali. Google se i tak snaží třídit odpovědi podle toho jak jsme zadávali(pořadí slov, slova v blízkosti u sebe) ale tyto přesné výrazy mnohdy velice vytřídí nalezené odpovědi.

+

Znak plusu by měl fungovat jako logický And. Avšak podle mých zkoumání je praticky jedno, používate-li mezerník nebo znak plus oddělující mezi sebou vyhledávané výrazy (zkuste si query slon kočka a slon+kočka).

-

Za znak mínus se dávají všechny výrazy, které nechceme aby byly zahrnuty ve vyhledávání. Je jedno jestli výrazy nebo koncovky souborů, prostě daný string se ve vyhledávání nesmí vyskytnout.

Hexaznaky

Google prakticky všechny speciální znaky z query vypustí. Což je pro nás v určitých chvílích jistá nevýhoda. Ovšem můžeme se pokusit tuto vlastnost obejít změnou znaků za hexa hodnoty. Tedy například znak / nahradíme za %2F. Bohužel toto funguje pouze u určitých url adres(těch, které interprety jako php). Nebo pokud chcete v url adrese použít mezeru, zadejte místo mezery %20 atd.

? *

Poslední věcí, kterou můžete ovlivňovat výsledky je dosazování znaků ? a *. Tyto znaky mají stejnou funkci jako při vyhledávání v na lokálním disku. Pro připomenutí znak * nahrazuje libovolný počet libovolných znaků a znak ? nahrazuje libovolný jeden znak.

-[Objasnění]-

Nyní se podíváme, co se vlastně dá takovým Google hackingem najít. Hledání by se dalo rozdělit do dvou možná třech podskupin. První skupina jsou data, od různých multimediálních souborů až po databáze s hesly. Druhá skupina jsou systémy, sem patří vyhledávání různých http serverů nebo webových aplikací. Třetí skupina jsou zařízení jako tiskárny, webové kamery, routery, tedy všechno co lze nastavovat přes webové rozhraní.

Tím se odstáváme k omezením, na google se (překvapivě) dají vyhledat věci, které jsou dostupné přes http protokol, ještě se dají (ve speciálních případech) nalézt nějaké ftp servery, ale tímto končí. Dále je velice důležité si uvědomit, jaké věci můžeme najít, respektive jaké věci je google schopen zaindexovat. Na všechny materiály, které můžete najít existuje hypertextový odkaz, url je někde v textu na stránce a nebo je daný prostor bez indexu a je vylistován. To ale neznamená, že všechny zaindexované soubory se nedostaly do google db náhodou, velice se dá totiž lehce přepsat nebo přehlídnout adresář, který může obsahovat citlivé informace. Ale pokud máte někde hluboko v adresářové struktuře nezaheslovanou složku _backup a nenavádíte na ní odkaz, ani adresáře nad ním nejsou přístupné/vylistované, tak google nemá šanci jentak takový adresář najít a zaindexovat. Google pracuje pouze s textem a odkazy, žádné speciální vyhledávací algoritmy v sobě nemá. Ale na druhou stranu, může u Vás na serveru někdo najít takovýto adresář nebo chybu, její popis vložit někam do fóra a to už k indexaci dojít může. A to vám můžu zaručit, že hodně lidí takovéto adresáře hledá.

No myslím, že objasněno je vše podstatné, takže se můžeme konečně začít zabývat tím, proč jsem tento dokument vlastně psal.

Poznámka: praxí získáte postupně větší zkušenosti a budete psát lepší a lepší queries a ty co uvedu níž rozhodně nepovažuji za dokonalé, rozhodně je na nich co vylepšovat a obměňovat, ale na to už doufám časem přijdete sami.

-[Vyhledávání dat]-

Začněme vyhledáváním dat, protože to je asi úplně první co budete chtít umět. Pokud jste zkoušeli moje queries u pokročilého vyhledávání, zjistili jste, že ačkoli názvy svádějí k získání zajímavých informací, z našeho pohledu získané výsledky jsou prakticky k ničemu. Jednotlivé queries musíme ještě malinko upravit.

Vyhledávání dat bych rozdělil do dvou základních skupin, na soubory, které lze vyhledávat podle obsahu a na ty, které nelze. Pustíme se nejdříve do těch, u kterých nelze, tedy víceméně multimediální soubory. Jako první věc si ukážeme vyhledávání souborů mp3, ogg, wma, ..., prostě hudebních. Na to existuje několik triků:

1) Pomocí "index of"

Jak jistě víte, naprostá většina web serverů sestavují stránky s obsahem daného adresáře tak, že v titulku stránky a první text na stránce je ve tvaru "index of nazev_adresare" následované výpisem všech souborů v daném adresáři. Proto zkusíme zadat do goolu něco jako intitle:"index of", chceme hledat mp3 soubory, takže můžeme přidat slovo "mp3" za "index of" nebo použít inurl"mp3" a ještě je dobré přidat slova jako "size" "name" "last modified" apod. A nakonec, pokud hledáme konkrétní skupinu přidáme její název na konec query. Výsledná query tedy může vypadat takto: "intitle:index inurl:/mp3 name size dead kennedys". Samozřejmě Možná ste si všimli, že jsem vynechal slovíčko "of", ale ve i když ho tam připíšete, tak se výsledek příliš nezmění, protože google takto krátká osamocená slovíčka ignoruje a intitle registruje pouze slovo "index".

2) Pomocí názvu

Postup je jednoduchý, napíšeme část songu, negujeme všechny stránky s html obsahem nebo přímo určíme filetype:mp3(ogg, ...) a případně doplníme ještě nějakým "inurl" parametrem. Ovšem tato metoda imho nepracuje úplně uspokojivě, mě osobně se nepodařilo tímto způsobem dosáhnout moc uspokojivých výsledků, ta první metoda je mnohem rychlejší a efektivnější.

3) Ftp hledání

Další možností je zkusit najít zaindexovaný ftp server s hudbou. Query složíme, že na začátku určíme protokol, tedy "inurl:ftp", pro jistotu odendáme www, a napíšeme co hledáme. Tedy výsledná query může vypadat takto: "inurl:ftp mp3 -www "parent directory" korn".

Tak to by k vyhledávání hudby mohlo stačit. Podobným způsobem lze vyhledávat i video soubory, ovšem slova jako music nebo mp3 nahradíte video, avi, divx, movies/z(starší 18 let mohou porno, xxx :)).

Další na pořadí jsou ebooky. O těch jen stručně, většina ebooku je ve formátu pdf(další pak ps,doc,...), takže budeme opět hojně využívat inurl a filetype. Samozřejmě těch queries může být celá řada, ovšem dominantní bude určitě "filetype:_typ_". Nebo pokud neznáte přímo koncovku souboru použijte tento řetězec "-html -htm -php -asp" ovšem potom query musíte upřesnit např.: nějakým inurl parametrem, jinak budete muset k dosažení cíle prolézat obrovské množství nepotřebných stránek. Chcete-li vidět jak lze tyto principy efektivně využít ve spojení s php, mlžete se podívat na http://library.biiter.com/ [5], kde máte vyhledávání ebooku zautomatizované.

Tímto jsme skončili s první skupinou a vrhneme se na skupinu druhou, tedy soubory s textovým obsahem. Ale to až v druhé části článku, která by měla vyjít asi tak za týden.

URL článku:
https://security-portal.cz/clanky/pokro%C4%8Dil%C3%A9-vyu%C5%BE%C3%ADv%C3%A1n%C3%AD-googlu-%C4%8D%C3%A1st-1

Odkazy:

[1] https://security-portal.cz/users/czokl
[2] https://security-portal.cz/category/tagy/programming
[3] https://security-portal.cz/category/tagy/security
[4] http://www.google.com/advanced_search
[5] http://library.biiter.com/