Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Ciência da Computação
Grupo de Resposta a Incidentes de Segurança

Rio de Janeiro, RJ – Brasil

Google: Ferramenta de ataque (e defesa) a sistemas

GRIS-2005-A-001

Breno Guimarães de Oliveira

A versão mais recente deste documento pode ser obtida na página oficial do GRIS

GRIS – Grupo de Resposta a Incidentes de Segurança
CCMN Bloco I 1° andar
Salas: I1021
Av. Brigadeiro Trompowski, s/nº
Cidade Universitária - Rio de Janeiro/RJ
CEP: 21949-900
Telefone: +55 (21) 2598-3309

Sumário

Introdução.....3
Como o Google funciona.....3
Inclusão e Exclusão.....3
Curingas.....4
Operadores avançados.....4
Usando o Google como ferramenta de ataque.....6
Ataque #1 – Identificação de servidores.....6
Ataque #2 – Servidores negligenciados.....7
Ataque #3 – Listagem de diretórios.....8
Ataque #4 – Senhas.....8
Ataque #5 – Explorando bancos de dados.....8
Ataque #6 – Explorando relatórios de segurança.....9
Ataque #7 – Usando o Google como um web proxy.....9
Ataque #8 – Fazendo o "googlebot" lançar ataques por você.....10
Usando o Google para se defender de ataques.....10
Consertando o estrago que já foi feito.....11
Bibliografia.....11

Introdução:

É raro encontrar alguém hoje em dia que use a Internet e não conheça a ferramenta de busca "Google". Trata-se de um fenômeno da Internet atual devido especialmente a sua rápida resposta, algoritmos sensatos de classificação de páginas e sua enorme base de dados. Mas para ter essa base de dados, o Google (e ferramentas de busca em geral) não se preocupa em discriminar paginas e arquivos públicos de dados sensíveis, contando apenas com a boa administração de cada servidor para impedir a inclusão de dados sensíveis nos resultados de buscas.

Como o Google funciona:

Existem atualmente duas principais vertentes de mecanismos de busca na Internet: as que catalogam páginas sob demanda e as que fazem uso de pequenos programas de varredura conhecidos como "agentes buscadores", "spiders" ("aranhas", em inglês), ou ainda "webcrawlers". O Google se encaixa nessa última vertente. Seu agente, o "googlebot", varre a Internet diariamente, adicionando e removendo páginas em sua base de dados e dando uma pontuação a cada uma, de acordo principalmente com a quantidade de referências feitas a ela por outras páginas. Assim, quanto mais referências independentes a ela uma página tiver, mais importante ela será considerada pelo Google, e mais perto do topo da lista de resultados ela estará. Isso garante que os primeiros valores retornados por sua busca sejam das páginas mais "bem cotadas" dentre todas as que contém as palavras que você procurou.

É muito fácil utilizar o Google e outras ferramentas de busca para encontrar o que se deseja na Internet: basta digitar uma ou outra palavra chave, apertar o botão de busca e esperar os resultados. O sistema retornará páginas que tenham TODAS as palavras fornecidas, dando preferência às mais bem classificadas, e àquelas cujas palavras chave fornecidas estejam mais próximas entre si. Para restringir sua busca, simplesmente adicione mais palavras. Para buscar por frases específicas, escreva a frase entre aspas duplas ("") e pronto. O que muita gente não sabe é que o Google oferece ainda uma série de recursos para pesquisas avançadas e refinamento de resultados, e são esses os recursos que veremos agora.

Inclusão e exclusão:

Como visto anteriormente, o Google retorna páginas que contenham todas as palavras fornecidas. No entanto, muitas palavras comuns (como "com", "a", etc.) são ignoradas por aumentarem consideravelmente o tempo de busca e raramente influenciarem nos resultados. Além disso, o Google *nunca* diferencia letras maiúsculas de minúsculas em suas buscas, e por padrão não considera acentuação. Para garantir que a palavra não seja descartada na busca, ou para garantir que ela será buscada com ou sem acentuação, exatamente como foi escrita (diferenciando "mao" de "mão", por exemplo, mas sem diferenciar "mão" de "Mão"), adicione o símbolo "+" imediatamente antes da palavra ou frase, sem espaços entre ela e o "+", mas com espaço entre cada palavra. É possivel ainda adicionar termos que você *não quer* que apareçam nos resultados através do símbolo "–", e buscar por páginas que contenham um ou outro termo, através da palavra reservada "OR", que deve ser sempre escrita em maiúsculas ou como o símbolo de barra vertical (|). Entrar com a busca, por exemplo,

```
+segurança  linux OR windows -"mac OS"
```

retornará páginas que contenham a palavra "segurança" escrita com acento, os termos "linux" ou "windows" (aceitando páginas que contenham pelo menos um desses termos), mas que não tenham a expressão "mac OS".

É possível ainda utilizar parêntesis para agrupar termos a serem incluidos ou excluidos na pesquisa.

Curingas:

O Google não aceita a utilização de símbolos especiais (também chamados de "curingas") como a interrogação (`?`) e o asterisco (`*`) para obter resultados variados. Pesquisar por "`seg*`", por exemplo, não fará o Google procurar por "`seguro`" ou "`segurança`".

No entanto, os caracteres ponto (`.`) e asterisco (`*`) tem uma função especial no Google. O primeiro é um substituto para um único caractere, e o segundo para uma única palavra. Se você quiser, por exemplo, buscar pelos termos

```
+uma*sensata
```

O Google vai retornar páginas que contenham frases como "`uma abordagem sensata`", "`uma medida sensata`", "`uma maneira sensata`", etc. Já o ponto indica que apenas um caractere separa as duas palavras, e é geralmente usado para indicar espaços ou caracteres especiais de escape. Note que não devem ser colocados espaços entre os termos e os curingas.

Operadores avançados:

Um dos recursos mais poderosos do Google e ao mesmo tempo desconhecido pela maioria dos usuários são os ditos "operadores avançados". Tais operadores ajudam a refinar sua busca de maneira impressionante e possuem a seguinte sintaxe:

```
operador:termo_de_busca
```

Note que não devem ser colocados espaços entre o operador, o "dois pontos" e o termo de busca. Os diferentes operadores são:

- SITE: busca por termos que apareçam em qualquer ponto do endereço, incluindo caminho completo da página.

- INURL: busca por termos contidos em qualquer ponto da URL, sem considerar o caminho relativo.

- FILETYPE: exibe apenas arquivos com a extensão especificada (.html, .php, .exe, .pdf, etc). O ponto (`.`) não precisa ser incluído.

- INTITLE: busca por termos que apareçam no título da página, ou seja, entre os comandos HTML `<title>` e `</title>`. Esse é o texto que costuma aparecer no topo da janela de seu navegador Web.

Veja o diagrama à seguir para melhor compreender o escopo de cada operador:

Outros operadores:

- INTEXT: Procura por termos contidos apenas no corpo de texto da página, sem considerar termos no título ou em links da mesma.

- NUMRANGE: Esse operador foi recentemente substituído por uma sintaxe mais simples. Agora, para procurar por páginas que contenham uma faixa qualquer de números, basta digitar PRIMEIRO..ÚLTIMO, onde PRIMEIRO é o número inicial, e ÚLTIMO é o final. Procurar, por exemplo, por "100..200" vai retornar todas as páginas que contenham pelo menos um número dentro dessa faixa.

- DATERANGE: busca apenas por páginas indexadas pelo Google em um intervalo entre duas datas diferentes. O valor da data, no entanto, deve ser fornecido na forma conhecida como "Data Juliana", um número inteiro que indica o número de dias passados desde 1 de janeiro de 4713 a.C. Um conversor instantâneo de datas pode ser encontrado em `http://www.24hourtranslations.co.uk/dates.htm`. Para utilizá-lo, modifique o dia, mês e ano desejado no campo "Greg. Cal.", e utilize o valor retornado na caixa "Julian Date". Até a data de publicação deste artigo esse operador não era completamente funcional e nem sempre retornava o esperado.

- INANCHOR: busca por termos contidos nos links da página, ou seja, na palavra ou frase exibida entre os comandos HTML `<a ...>` e `</a>` das páginas.

- LINK: exibe páginas que contenham links para uma determinada URL, que deve ser passada como parâmetro para essa opção. Essa opção é extremamente útil para saber que páginas referenciam um determinado sítio ou página web.

- CACHE: exibe a versão armazenada pelo Google de uma determinada URL, que deve ser passada como parâmetro para essa opção.

- RELATED: busca por páginas similares (na opinião dos algoritmos de busca do Google) a página indicada na URL que deve ser passada como parâmetro para essa opção.

- INFO: exibe algumas informações que o Google possui sobre a URL indicada, que deve ser passada como parâmetro para essa opção.

Alguns operadores admitem o prefixo "**ALL**", para indicar que todas as palavras a seguir devem ser incluídas na busca daquele operador. Por exemplo, buscar por "`ALLINTITLE:gama beta alfa`" (sem as aspas) vai retornar as páginas cujo título possui os três termos, independentemente da ordem em que aparecem.

Até a data de publicação deste artigo existiam ainda os operadores "**PHONEBOOK**" para pesquisar telefones (essencialmente dentro dos EUA), "**STOCKS**", que busca pela cotação dos termos passados na página de finanças do provedor de serviços "Yahoo", e "**DEFINE**", para buscar por definições de termos e expressões. O Google possui ainda uma calculadora embutida. Para usá-la, basta digitar uma expressão matemática e aguardar a resposta. A calculadora também entende constantes e faz conversões entre bases, pesos e medidas. O uso dessas e outras características do Google não entra no escopo deste artigo e, portanto, não será abordado.

Usando o Google como ferramenta de ataque:

A utilização de agentes buscadores para catalogar páginas, ao mesmo tempo em que aumenta consideravelmente o número de respostas da ferramenta, traz o problema de muitas vezes catalogar mais do que os administradores de páginas gostariam. Isso porque muitos servidores conectados a Internet – e, portanto, passíveis de serem catalogados por agentes buscadores – não foram configurados de forma segura, contando apenas com o fato de não serem conhecidos para se protegerem. Assim, se os agentes do Google encontrarem, por exemplo, arquivos sigilosos ao fazerem uma varredura nos servidores de sua empresa, eles irão adicionar tais arquivos e páginas à base de dados do Google na primeira oportunidade, já que não são capazes de diferenciar o que é público do que é confidencial. Dessa forma, muita informação "sensível" pode ser encontrada, bastando que se saiba *o que* procurar. Não bastasse isso, o Google armazena no "cache" uma versão da sua página ou arquivo, de modo que mesmo corrigindo o problema do seu servidor, seus dados podem ainda estar expostos. Vejamos alguns tipos de ataques que podem ser feitos através do Google.

Nota: é importante observar que a maioria dos ataques e técnicas aqui exemplificados não são nenhuma novidade. Existem registros sobre alguns destes que datam de 2001, o que indica que usuários maliciosos vêm explorando tais características de sistemas de busca a mais tempo ainda.

Ataque #1 – Identificação de servidores

Para identificar todas as páginas da Google que o próprio Google já catalogou, por exemplo, basta digitar

```
inurl:google.com
```

Para encontrar servidores extras dentro do mesmo domínio, subtraia o domínio principal da busca anterior. No exemplo, podemos entrar com

```
inurl:google.com –inurl:www.google.com
```

para listar diversos subdomínios dentro de "google.com" (como, por exemplo, "images.google.com", "gmail.google.com", "desktop.google.com", etc).

É possível ainda encontrar serviços remotos em execução. Veja os exemplos à seguir:

```
intitle:"VNC Desktop " inurl:5800
intitle:"remote desktop web connection"
intitle:"Terminal Services Web Connection"
allintitle:"microsoft outlook web access" logon
```

O primeiro procura por servidores "VNC" em execução. O segundo e o terceiro referem-se, respectivamente, à Área de Trabalho Remota ("Windows Remote Desktop") e aos serviços de terminal ("terminal services") da Microsoft. O último busca por páginas de acesso remoto ao *webmail* Microsoft Outlook.

Ataque #2 – Servidores negligenciados

Muitos servidores de páginas web, como o Apache, exibem páginas padrão ao serem instalados. Isso é feito essencialmente para mostrar ao administrador que o serviço está sendo executado sem problemas, e oferecer dicas sobre como proceder em seguida. Acontece que muitas pessoas acabam instalando servidores sem saber, e embora a existência da página padrão não indique especificamente uma vulnerabilidade, costuma mostrar que o servidor em questão foi negligenciado pelos administradores, e possivelmente está com suas configurações padrão. Afinal, se o administrador não se deu ao trabalho de sequer modificar a página principal do servidor, é muito provável que ele também não tenha se preocupado em atualizar o mesmo, ou em tornar o próprio sistema mais seguro. Buscar por

```
intitle:"test page for Apache"
```

ou,

```
intitle:"Página teste para a instalação do Apache"
```

retorna uma lista de servidores que estão exibindo a página padrão do Apache. Para procurar por páginas padrão do IIS, poderíamos buscar por

```
intitle:welcome.to.IIS.4.0
```

ou variações como

```
intitle:"welcome to windows 2000 internet services"
intitle:"welcome to windows xp server internet services"
"under construction" "does not currently have"
```

para servidores mais novos.

Ataque #3 – Listagem de diretórios

Configurar servidores adequadamente pode ser uma tarefa bastante confusa, e de fato muitos administradores cometem erros na hora de determinar que diretórios do sistema podem ou não ser acessados via Internet. Além disso, servidores costumam permitir a listagem dos arquivos e subdiretórios dentro de um diretório que possa ser acessado mas não tenha a página de nome padrão (geralmente "index.html", embora isso possa ser modificado nas configurações do servidor). Isso permite que usuários maliciosos procurem por listagem de diretórios sensíveis em seu servidor, ou ainda por arquivos específicos dentro de seu sistema, que podem ser acessados pelo nome mesmo que seu servidor esteja configurado para não listar o conteúdo de diretórios (desde que, é claro, o servidor tenha permitido acesso ao arquivo devido a má configuração do servidor). É possível, por exemplo, buscar por

```
intitle:index of/admin
```

retorna listagem de diretórios chamados "admin", comumente utilizados por administradores de sistemas para guardar arquivos e dados – possivelmente sensíveis. Buscar por arquivos e subdiretórios específicos dentro de listagens diretórios também é possível, como por exemplo:

```
intitle:"index of" .htpasswd
intitle:"index of" backup
```

Ataque #4 – Senhas

Muitos servidores mal configurados tornam públicos seus arquivos de registro ("logs"), permitindo assim que usuários maliciosos obtenham senhas de sistemas (muitas vezes com privilégios de administrador) sem trabalho algum, bastando buscar por

```
filetype:log inurl:"password.log"
```

Ataque #5 – Explorando bancos de dados

Grande parte dos servidores web precisa de serviços de banco de dados instalados, mas muitos não são bem configurados e acabam fornecendo informações sigilosas, como tabelas inteiras, que podem conter até mesmo campos com nome de usuários e senhas válidas dentro do sistema. A seguinte busca procura por tais tabelas:

```
"# dumping data for table" (username|user|users) password
```

Conhecendo um pouco da estrutura de arquivos de uma base de dados SQL, é possível ir diretamente atrás de arquivos que contenham senhas, como por exemplo através da busca:

```
filetype:properties inurl:db intext:password
```

É possível ainda identificar bancos de dados vulneráveis a ataques de "injeção de SQL", ao pesquisarmos por mensagens de erro que tipicamente denunciam esse problema:

"ORA-00921: unexpected end of SQL command"
"ORA-00933: SQL command not properly ended"
"unclosed quotation mark before the character string"

Ataque #6 – Explorando relatórios de segurança

Administradores preocupados com a segurança de seus sistemas costumam executar programas específicos que realizam varreduras de segurança e identificam vulnerabilidades em seus servidores. Procurar por

"This file was generated by Nessus"
ou
"This report lists" "identified by Internet Scanner"

retorna desde páginas exemplo até relatórios reais, que indicam as vulnerabilidades encontradas em determinados servidores, que colocaram os relatórios das ferramentas (como o "nessus" ou o "ISS", do exemplo acima) em uma área pública.

Ataque #7 – Usando o Google como um web proxy

A possibilidade de se traduzir páginas é um dos grandes atrativos do Google. Através da página

http://translate.google.com/translate_t

podemos digitar uma URL qualquer e o Google fará a tradução da página de acordo com o idioma desejado. O procedimento é simples: O Google acessa a página, traduz, e retorna ela para você. Na prática, você não fez nenhuma conexão direta à página desejada, e o Google atuou como um web proxy para você. O único problema desse procedimento é que você precisaria traduzir a página desejada de algum idioma para outro, e visualizá-la somente no idioma destino. No entanto, isso pode ser facilmente contornado. A sintaxe retornada pelo Google para as traduções é no seguinte formato:

http://translate.google.com/translate?u=PAGINA&langpair=LANG1|LANG2

onde `PAGINA` é a URL completa desejada, `LANG1` é o código para o idioma original da página, e `LANG2` é o código para o idioma destino. Manipulando esses valores, podemos colocar o mesmo idioma como origem e destino, e assim ver a página em seu idioma original, utilizando o Google como web proxy. Para ver o conteúdo da página do GRIS (em português) através do Google, basta digitar

http://translate.google.com/translate?u=http://www.gris.dcc.ufrj.br
&langpair=pt|pt

Idiomas identificados pelo Google até a data de publicação deste documento são:

```
de (alemão)
es (espanhol)
fr (francês)
it (italiano)
pt (português)
us (inglês)
```

Ataque #8 – Fazendo o "googlebot" lançar ataques por você

Como visto no início deste documento, o "googlebot" é o agente responsável pela busca e classificação das páginas dentro de servidores. Pedir ao agente para visitar sua página é um procedimento fácil, e que pode ser feito de muitas maneiras diferentes. Uma vez dentro de sua página, o "googlebot" (e qualquer outro agente de serviços de busca, na verdade) vai registrar e seguir cada um dos links que você incluir dentro da mesma, não importa quais sejam. Um usuário pode, portanto, adicionar links de natureza maliciosa em sua página e apenas esperar os agentes surgirem para fazer o "trabalho sujo" por ele. Os agentes buscadores se encarregarão de executar o ataque e, não bastasse isso, ainda podem colocar os resultados devolvidos pelas vítimas (caso existam) publicamente em suas páginas de busca, para que qualquer um (inclusive o atacante) possa ver. Exemplos links com ataques potenciais incluem:

```
http://algumhost/cgi-bin/script.pl?p1=`ataque`
http://algumhost:54321/ataque?`id`
http://algumhost/AAAAAAAAAAAAAAAAAAAAAAAAAAAA...
```

Repare que é possível ainda manipular a sintaxe das URLs para mandar o agente acessar o servidor alvo em portas específicas (na segunda linha de exemplo, mandamos ele acessar a porta 54321).

Existem muitas outras formas de ataque possíveis através de mecanismos de busca. Sabendo o que procurar, é possível utilizar o Google para encontrar informações que vão de dados de um servidor até senhas de banco e números de cartão de crédito.

Usando o Google para se defender de ataques:

Da mesma forma que o Google pode ser usado por usuários maliciosos a fim de atacar seus servidores e clientes, você também pode utilizá-lo para proteger os mesmos através de algumas simples ações e constante vigilância. Essencialmente, não coloque informações sensíveis em seus servidores, ainda que temporariamente. Configure seus servidores com atenção redobrada e verifique regularmente sua presença (e a presença de seu sistema) dentro do Google, utilizando o operador "site:" para fazer pesquisas em todos os seus servidores. Este operador também aceita números de IP como parâmetro, então é possível utilizá-lo em seu servidor de IP fixo mesmo que ele não tenha um nome de domínio válido.

O procedimento de verificação acima pode ser automatizado através da utilização de ferramentas gratuitas disponibilizadas na Internet, dentre elas o "sitedigger", da Foundstone, que pode ser obtido em:

http://www.foundstone.com/resources/proddesc/sitedigger.htm

e o Wikto, da Sensepost, que pode ser obtido em:

http://www.sensepost.com/research/wikto/

Consertando o estrago que já foi feito:

Ainda que você tome as devidas atitudes ao identificar um problema em seus sistemas através do Google, este pode continuar exibindo sua página, dados ou arquivos indesejados através do sistema de armazenamento de páginas. Para resolver esse problema, basta avisar ao Google que você quer a referência anterior à sua página atualizada ou removida, acessando

http://www.google.com/webmasters/

e seguindo as indicações. Naturalmente, para remover a referência a um servidor, você precisará provar que é o administrador do mesmo.

Bibliografia:

Página oficial do Google
URL: http://www.google.com

Long, J. – The Google Hacker's Guide
URL: http://johnny.ihackstuff.com/modules.php?op=modload&name=Downloads&file=index&req=viewdownload&cid=3

Long, J. – You Found That on Google? Gaining awareness about "Google Hackers". Palestra ministrada durante o evento Blackhat/Defcon 2004.
URL: http://johnny.ihackstuff.com/security/premium/BH2004FINAL.htm

Zalewski, M. – Against the System: Rise of the Robots
Artigo publicado no zine Phrack número 57
URL: http://www.phrack.org/phrack/57/p57-0x0a

GooFresh Date-based Syntax Definition
URL: http://www.researchbuzz.org/archives/001405.shtml