Índice das dicas

Baixando sites inteiros com o wget

Criado 17/out/2007 às 12h38 por Marcos Elias Picão 7 4

Muitas vezes você encontra um site cheio de coisas interessantes, mas não dispõe de tempo para ler on line na hora. Ou, encontra algo tão importante que pretende guardar em CD ou DVD, para nunca perder. A natureza dos sites da web não permitem isso. Você entra, acessa um site e navega pela página. Se quiser outro conteúdo, deve clicar num link, então o mesmo é descarregado até a sua tela. Fica complicado, por exemplo, querer salvar um site inteiro no seu computador clicando nos links e salvando página por página.

Lá vem a computação novamente, com sua rapidez e "inteligência". Sim, inteligencia entre aspas porque as máquinas em si são burras, meros papagaios, como você sabe só fazem aquilo que foi previamente programado. Há programas que podem baixar os sites inteiros, sem que você precise entrar e clicar em link por link. O que esses programas fazem é seguir os links das páginas, deifinidos em HTML pela tag <a href=...>. No final do processo, você pode ter todo o site ou várias páginas e arquivos do seu interesse.

Existem vários programas que se propõem a isso, a maioria para Windows, normalmente pagos. Quase sempre o que se encontra são sharewares, que se limitam a funcionar por um período determinado ou baixam apenas um certo número de páginas, sendo liberados somente na versão completa. Há um gratuito e open source, o HTTrack (WinHTTrack, na versão para Windows), é bom e relativamente bastante usado, por ser fácil e bastante personalizável. Mas há muito tempo há uma outra solução gratuita e aberta, muito conhecida pelos usuários de Linux: o **wget**. O wget é um programa criado inicialmente para o ambiente Unix/Linux, cujo objetivo principal é baixar arquivos da internet. Ele pode ser usado em scripts, tornando a programação de diversas "aplicações" bem mais fácil. Para felicidade de quem usa Windows, saiba que há uma versão portada dele para o Windows, que funciona exatamente da mesma forma da versão Unix. E para quem quer baixar sites inteiros ou várias páginas, ele também tem esse recurso. De quebra, é um programa bem pequenininho, operado via linha de comando, o que permite o uso fácil em scripts ou a criação de interfaces.

Antes de começar, tenha em mente que baixar sites inteiros pode prejudicar os sites. Além dos sites em si, essa tarefa que pode consumir muita banda de download deles, você pode acabar prejudicando outros sites que estejam hospedados no mesmo servidor, especialmente se for um hospedeiro compartilhado - mesmo comercial. Muita gente não tem noção de que baixar sites de uma vez ou acessar excessivamente arquivos grandes pode ser prejudicial, e faz isso sem saber; é bom que você esteja consciente para não abusar.

Obtendo o wget

Quase toda distribuição de Linux inclui ele. Caso não inclua, experimente instalá-lo usando o gerenciador de pacotes da sua distribuição. Se você usa Windows, pode baixá-lo em:

http://pages.interlog.com/~tcharron/wgetwin.html

O básico do wget

Para baixar um arquivo com o wget, basta dar o comando *wget* seguido pelo caminho do arquivo

na internet. Exemplo:

```
wget http://www.umsitequalquer.com.etc/arquivo.zip
```

E o arquivo será baixado na pasta atual do prompt de comando. Caso o download seja interrompido (porque o computador trave, ou o usuário tecle CTRL + C no prompt), é possível recontinuar do ponto em que parou, caso o servidor suporte o recurso. Basta informar, antes da URL, o parâmetro *-c*, e claro, usar a mesma pasta que contém o "pedaço" anterior do arquivo:

```
wget -c http://www.umsitequalquer.com.etc/arquivo.zip
```

Muita gente usa o wget como gerenciador de downloads (no Linux), em vez de baixar os arquivos pelos navegadores. Ele não acelera os downloads (como fazem os "aceleradores de downloads", que abrem várias conexões), mas permitir recontinuar do ponto em que parou é muito útil, especialmente para quem acessa com conexão lenta ou instável, que cai toda hora.

Baixando sites inteiros

Passando o parâmetro *-r*, ele baixará todos os arquivos encontrados no domínio. Cuidado, use com atenção e responsabilidade! Além de poder atrapalhar o site, você poderá baixar muita coisa inútil. A sintaxe seria:

```
wget -r http://www.umsitequalquer.com.etc
```

Ele seguirá os links definidos em HTML, pela tag <a href...> nas páginas. Links em JavaScript ou em flash não serão reconhecidos, e as páginas não serão baixadas. No caso do JavaScript, é quase impossível um programa sair baixando tudo realmente, pois há várias formas de se "programar" links em JavaScript.

Por padrão, o wget ignora os arquivos que o produtor do site pediu para ignorar, por meio de um arquivo especial, o "robots.txt". Esse arquivo serve para os motores de busca de sites de pesquisa; eles lêem o arquivo e ficam sabendo para quais arquivos não devem seguir os links. Isso impede que o wget baixe determinados arquivos, arquivos esses que muitas vezes são justamente os que você precisa. Para fazê-lo ignorar os arquivos "robots.txt" e baixar tudo o que encontrar, basta usar o parâmetro *-erobots=off*. Ficaria assim:

```
wget -r -erobots=off http://www.umsitequalquer.com.etc
```

Apesar de tecnicamente possível, é um meio não muito legal fazer isso. Essa opção *-erobots=off* não é comentada em muitos lugares. Novamente alerto, use com responsabilidade ou quando precisar "mesmo", afinal não devemos ser contra a divulgação de informações. Você pode usá-la, por exemplo, quando quiser baixar um arquivo linkado mas que não esteja disponível para ser indexado pelos buscadores (via robots.txt), já que o wget o ignoraria.

Uma dica para reduzir a quantidade de downloads é baixar apenas arquivos de determinado tipo, por exemplo, apenas páginas HTML. Isso é possível com o parâmetro *-A*, especificando a seguir o tipo de arquivo pela extensão. Veja:

```
wget -r -A ".html" http://www.umsitequalquer.com.etc
```

As páginas ASP, PHP, entre outras, serão convertidas em HTML pelo wget (até porque serão páginas já processadas quando o wget acessá-las), ou seja, você não precisará se preocupar em ficar digitando todas as extensões possíveis para as páginas. Da mesma forma, fica fácil baixar todas as imagens:

```
wget -r -A ".gif" http://www.umsitequalquer.com.etc
```

ou

```
wget -r -A ".jpg" http://www.umsitequalquer.com.etc
```

Depende do que você precisa.

Dica direta para baixar sites inteiros: o wget possui o parâmetro *-m*, ideal para fazer espelhamentos de sites (mirroring), onde ele já ativa as opções necessárias. Se seu objetivo é fazer um *mirror* do site, pode ir direto ao ponto:

· Jogo falso vendido no
· Linux Mint eliminando
· Desmontar KDE

23/03

· YouTube m estabilidade
· Empresas contratar; F
· Lançada v para o Nokia
· Blu-ray: O

22/03

· TP-LINK T
· Usando o
· Photoshop movimentaç

Notícias d

Artigos

Livr

Comp

```
wget -m http://www.umsitequalquer.com.etc
```

Outra coisa é a identidade do navegador. Alguns sites só oferecem o conteúdo ao verificarem que é determinado navegador que está acessando a página (como o IE, por exemplo), e o wget não conseguiria acesso aos arquivos. Pode-se fingir a identidade de navegador com esse parâmetro:

```
--user-agent="Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"
```

Colocando no texto entre aspas a identificação do navegador e sistema que você quiser.

Páginas protegidas com senha

Páginas que exibem uma tela de login personalizada, via web mesmo, provavelmente não poderão ser baixadas. Digamos que você queira baixar todo o conteúdo do Orkut, ou de uma comunidade ou fórum restrito. Ao acessar a página, vem uma página do site que pede o login e a senha. O wget pararia nela e se limitaria a pegar as páginas linkadas à ela diretamente.

No entanto, páginas que pedem senha diretamente pelo servidor, podem ser baixadas. Estas podem ser acessadas passando o nome e a senha diretamente na URL, com o wget não seria diferente:

```
wget http://usuario:senha@servidor.com.etc/arquivo.etc
```

Também pode ser usado para servidores FTP, bastando trocar o http:// por ftp://.

Outros parâmetros importantes

Através de diversos parâmetros podemos controlar melhor o donwnload dos arquivos. Eis alguns importantes:

```
-t0
```

Número de vezes que ele deve tentar baixar um arquivo, caso não consiga. Definindo zero deixa configurado como "ilimitado", o que faz ele ficar tentando até conseguir baixar. Observe que deve-se deixar o número logo ao lado da letra "t", sem espaço. É bom usá-lo mas com cuidado, num site com muitos arquivos e uma conexão ruim, você pode ficar forçando o download de vários arquivos que nunca serão baixados, à toa. Por precaução, defina um número menor, como -t3 ou -t5.

```
-c
```

Permite continuar downloads interrompidos, sem que ele precise baixar o que já foi baixado, incluindo pedaços de arquivos (e não apenas páginas). É bom usá-lo, principalmente quando não der tempo de baixar o que você quer em uma única vez.

```
-np
```

Ideal para baixar páginas de uma mesma área de um site. Com o -np (no-parent) ele baixa apenas os arquivos da URL definida sem pegar a pasta pai dela. Por exemplo, utilize caso queira baixar o site http://www.site.com.etc/algumacoisa, mas não o conteúdo do http://www.site.com.etc (e sim apenas da pasta /algumacoisa).

```
-T30
```

Define o timeout (limite de tempo) para 30 segundos. Quando ele fica à espera do arquivo, em conexões lentas, esse parâmetro orienta a refazer a conexão para puxar tal arquivo a cada 30 segundos, até conseguir. Observe a diferença que este é em letra maiúscula.

E os donos dos sites, como ficam?

Certamente não gostam nada nada disso. É possível aplicar alguns bloqueios diretamente no servidor (o que foge ao objetivo deste texto, além do que variará muito dependendo do software de servidor utilizado), por exemplo, bloqueando IPs que acessem várias páginas por segundo (mais precisamente um número maior de bytes definido por você), reduzir a velocidade, ou até mesmo banir o IP por um período.

Mas não haverá muito o que fazer para páginas em si. O pessoal dá um jeito, afinal se pode ser

acessado pelo navegador, poderá ser baixado e salvo.

Uma dica para evitar downloads automático de arquivos grandes, poupando banda do servidor, é usar um link definido em JavaScript. Praticamente nenhum programa copiador de sites identificaria o link, mas nos navegadores funcionaria normal. Estar com o JavaScript habilitado nos micros clientes hoje é praticamente *obrigatório*, não existe mais aquele papo de problemas de compatibilidade. Sendo assim, não se preocupe com o link em JavaScript, pois ele funcionará.

A criação do link fica a seu critério, evite colocar o nome do arquivo por inteiro. Um exemplo prático: definindo uma função na página:

```
<script language=javascript>
function BaixaCoisa(arq){
self.location.href = arq;
}
</script>
```

E chamando a função nos links assim:

```
<a href="javascript:BaixaCoisa('/arquivo.iso');">Clique aqui para baixar
o CD</a>
```

Se preferir, pode complicar mais, mas dará mais "trabalho" para implementar. Exemplo (este coloque diretamente no local onde for ficar o link):

```
<script language=javascript>
document.write('<a href=');
document.write('arquivo.zip');
document.write('>');
document.write('clique aqui');>
document.write('</a>');
</script>
```

Um outro meio um pouco mais complicado de implementar, é aplicar CAPTHA, aquela verificação de letras e números aleatórios em determinadas seções do site. O programa pára ali. Mas cuidado, a maioria dos sistemas não permitem o uso por deficientes visuais, o que prejudica legal a acessibilidade do seu site, especialmente se usar isso em áreas essenciais.

É isso. Responsabilidade sempre!

Por Marcos Elias Picão. Revisado 2/mar/2011 às 14h56 7 comentários

Comentários

Entrar e fazer comentário

wget
Criado 26/ago/2011 às 21h13 por **Leogh (anônimo)**

O link precisa ser um link direto?

WGET
Criado 22/jan/2011 às 11h55 por **Tomé (anônimo)**

Viva Marcos,

Trata de um assunto de interesse da comunidade free. Dou-lhe os parabéns por estas linhas que escreveu.. No entanto devemos todos, como o Marcos disse "Novamente alerto, use com responsabilidade ou quando precisar.." :)

wget
Criado 24/dez/2010 às 15h18 por **Ricardo (anônimo)**

Caraca!!! eita comando!!! baixa tudo mesmo!!!

wget
Criado 23/dez/2010 às 22h13 por **Gilvan Ritter (anônimo)**

Pra quem quer esperimentar recomendo baixar uma mina de ouro: http://serverapostilando.com/tutorials/

wget -m -erobots=off http://serverapostilando.com/tutorials/ /sua_pasta_aqui

baixe
Criado 8/set/2010 às 14h14 por **julio (anônimo)**

muito bom o artigo. já usei as dicas para baixar alguns manuais da net.

Obrigado.

Expandir réplicas

Num tem jeito *por* ***Adriano (anônimo)***

Eis tudo o que eu queria!
Criado 23/set/2010 às 02h59 por **Sharrukin (anônimo)**

Enfim fizeram um artigo com tudo o que eu precisava! Valeu mesmo!!!

Destaques

· 39 aplicativos indispensáveis para o Android
· Configurando rapidamente uma rede entre dois micros
· Como colocar legendas em vídeos
· Hackeando as senhas no Windows XP
· Crimpando cabos de rede
· Configurando a rede no Windows (atualizado)
· Qual a diferença entre notebook e netbook?
· Prompt de Comando do Windows
· Planos de dados: usando o celular como modem
· Celulares chineses
· Instalando o Apache + PHP + MySQL no Windows
· Limpando os arquivos temporários do Windows
· *Guia do Hardware agora é Hardware.com.br*
· **Hardware II, o Guia Definitivo**