



Web Robots

- Populares de 1993 a 1995
- Principais Aplicações
 - análise estatística
 - manutenção de sites
 - mirroring
 - descoberta de recursos
 - usos combinados
- Grande Número de Web Robots mal configurados
 - causaram grande número de aborrecimentos
- Análise de Custos e Benefícios de se utilizar Web Robots



Web Robots

■ Recursos de Rede e Carga dos Servidores

- uso considerável de bandwidth de rede
 - | robots operam continuamente durante períodos prolongados de tempo e muitas vezes fazem diversas buscas em paralelo
 - | mesmo partes remotas da rede podem sentir o "esgotamento de recursos" quando um robot está em operação
 - | uso da rede pode dar a impressão de ser "free", mas a medida que a demanda cresce, existe uma degradação sensível da qualidade de serviço
- demanda considerável de serviço de servidores
 - | dependendo de como um robot acessa as páginas de um servidor, pode haver uma carga considerável sobre ele, evitando que outros clientes possam ter acesso a seus serviços temporariamente
- evitar o uso de acessos do tipo "rapid fire"



Web Robots

■ Custo de Update dos Dados

- não há controle sobre a mudança, remoção e modificação de URLs na Web
- o protocolo HTTP provê o mecanismo "If-Modified-Since"
 - | este recurso somente pode ser utilizado se o robot mantém os dados sobre a última incursão sobre a página

■ Implementações Deficientes

- distribuição inconsequente de código
- erros comuns
 - | falta da história dos sites já visitados (ocasionando repetição)
 - | URLs sintaticamente equivalentes
 - | download de arquivos não-HTML, tais como GIFs e Postscript
 - | incompetência em lidar com sites com scripts (páginas dinâmicas)
 - buracos negros



Web Robots

- Possíveis ações de contorno
 - Determinação de regras para inclusão/exclusão de sites
 - Sumarização de Documentos
 - | uso de meta-informações contidas no texto (não há standard)
 - Classificação de Documentos
 - Determinação da Estrutura do Documento
 - | descoberta do valor de palavras dentro do documento
- Necessidade de Ética
 - autor de um web robot deve balancear seu desejo por informações com o impacto que o mesmo pode causar
 - justifica-se a operação de um web-robot, diante do custo que este pode causar aos outros ?



Web Robots

■ Guidelines for Robot Writers

- <http://info.webcrawler.com/mak/projects/robots/guidelines.html>

■ Reconsiderar

- precisamos realmente deste robot ?
- Há outras maneiras de se obter o mesmo serviço ?
- Existe outro robot, já em operação, que pode me prover a informação desejada ?

■ Ser Rastreável

- identificar seu robot ao Web Server (User-Agent field)
- identificar o autor do robot (e-mail, usando From field)
- anunciar o robot em listas de discussões (e.g. comp.infosystems.www.providers)
- anunciar seu interesse em acessar sites sistematicamente



Web Robots

■ Ser Rastreável

- seja informativo (use o field Referer para dizer por que está acessando o site)
- esteja comunicável enquanto o robot está "on-line"
 - ┆ suspenda o robot, quando não estiver disponível
- notifique o administrador do seu sistema sobre seus planos de rodar um web robot

■ Testar Localmente

- utilize os sites locais para realizar testes
- analize o desempenho do robot e estime como o mesmo seria escalável para dimensões maiores de URLs vasculhadas



Web Robots

■ Não Desperdice Recursos

- faça o robot rodar devagar (faça-o dormir de quando em quando)
- rotacione os servidores acessados (evite o fire-access)
- use If-modified-since e HEAD sempre que possível ao invés de GETs
- peça somente o que deseja (use o campo Accept)
- evite que sub-referências à mesma página sejam acessadas como uma nova página
- mantenha uma lista de lugares que não deve visitar
- cheque se as URLs são válidas - cuidado com o / ao final de diretórios e com a concatenação de nomes de sub-URLs
- cheque os resultados devolvidos pelo servidor



Web Robots

■ Não Desperdice Recursos

- Verifique se uma determinada URL já não foi visitada - evite loops ou repetições
- verifique se duas URLs distintas não correspondem à mesma página
- rode o robot em horários oportunos
- demore para revisitar um mesmo site
- faça uma lista de links voláteis (tais como "what's new") - utilize-os para a obtenção de novos links
- paralise o robot, de tempos em tempos
- não faça queries, ou preencha formulários - abandone os links que contém formulários



Web Robots

- Acompanhe o funcionamento do Robot
 - faça logs de tudo que é efetuado pelo robot
 - utilize estes logs para detectar mal-funcionamento do robot
 - faça o robot ser interativo - proveja meios para suspender ou paralisar o robot, quando necessário, sem ter que abortá-lo
 - esteja preparado para reclamações
 - seja compreensivo - instrua o robot para não visitar sites onde tenha havido reclamações
 - não tente violar barreiras para o acesso de robots
 - respeite o desejo dos administradores dos servidores
 - estime o espaço útil que o robot pode consumir em sua máquina, de modo que o mesmo tenha consciência se está próximo de seu limite



Web Robots

■ Compartilhe os resultados

- mantenha os resultados da atividade do robot - faça com que o robot possa utilizá-los quando re-iniciar suas atividades
- disponibilize esses resultados para o acesso de terceiros
- publique os resultados obtidos - anuncie em listas de discussões
- reporte os erros encontrados durante a execução do robot e notifique as ações determinadas para corrigí-los

■ Atualmente

- maioria das implementações de robots seguem estas recomendações
- redução do número de ocorrências com incidentes negativos



Padrões para a Exclusão de Robots (SRE)

■ Publicado em Junho de 1994

- <http://info.webcrawler.com/mak/projects/robots/norobots.html>
- lista de discussões sobre web-robots
- norma não-oficial e não-comercial
- guideline para a orientação de desenvolvedores de robots

■ Método

- URL “/robots.txt” localizado na raiz do servidor
- conteúdo do arquivo indica a política desejada do servidor em relação a robôs.
- Este arquivo pode ser construído automaticamente a partir de outros localizados nos diretórios particulares que são originados a partir da raiz.



Padrões para a Exclusão de Robots (SRE)

■ Formato

- linhas construídas da seguinte forma:

"<field>:<optionalspace><value><optionalspace>"

■ Fields

■ User-agent

- | declara o nome do robot para o qual se está especificando a política de acesso
- | nomes são case-insensitive
- | um valor * indica a política default para robots não-nominados

■ Disallow

- | o valor deste campo indica uma URL parcial que não deve ser visitada
- | valores vazios indicam que todas as URLs podem ser acessadas



Padrões para a Exclusão de Robots (SRE)

■ Exemplos

- nenhum robot deve visitar nenhuma URL começando com `"/cyberworld/map/"`, ou `"/tmp/"` ou `"/foo.html"`:

```
User-agent: *  
Disallow: /cyberworld/map/ # This is an infinite virtual URL space  
Disallow: /tmp/ # these will soon disappear  
Disallow: /foo.html
```

- exceção para o robot cyberrmapper:

```
User-agent: *  
Disallow: /cyberworld/map/ # This is an infinite virtual URL space  
# Cybermapper knows where to go.  
User-agent: cyberrmapper  
Disallow:
```

- nenhum robot permitido

```
# go away  
User-agent: *  
Disallow: /
```



Avaliação do SRE

■ Áreas que se desejam excluir

- espaços de URL onde robots podem ser confundidos (black holes)
- espaços de URL de recursos intensivos (e.g. páginas dinâmicas)
- documentos que atraem um tráfego muito intenso (e.g. material erótico)
- documentos que possam representar um site de maneira negativa (e.g. arquivos de bugs e bug-reports)
- documentos que não são úteis para acesso geral (e.g. informação local)

■ Arquitetura do SRE

- simples de administrar, de implementar e de distribuir
- permite a distribuição do controle de acesso a páginas particulares



Avaliação do SRE

■ Problemas Operacionais

- problemas relacionados à administração do SRE
- acesso e/ou montagem do arquivo /robots.txt
- impossibilidade de generalizar tipos de arquivos em um diretório
- redundância para robôs específicos
- escalabilidade (dificuldade com múltiplas páginas)

■ Problemas Relacionados à Web

- nomes de domínio errados
 - ┆ trechos do site podem obter nomes independentes
- mirrors
 - ┆ múltiplas máquinas com o mesmo conteúdo
- Updates
 - ┆ robots não sabem quando a página foi alterada



Avaliação do SRE

- Elementos ainda não contemplados
 - múltiplos prefixos por linha
 - tempo entre múltiplos acessos
 - frequência de re-visita - revisita ao arquivo robots.txt
 - horas preferenciais para visita
 - permissão de visita mas não de uso de links internos
- Extensões além do SRE
 - sugestões de URLs a se visitar
 - meta-dados com informações locais
 - dados de contato do administrador do site
 - descrição do site
 - informações geográficas sobre o site



Web Robots e Meta-Dados

■ Meta-Dados

- Dados sobre os dados
- Informações contidas em um documento web, trazendo informações sobre o documento e/ou sobre partes e trechos do documento
- facilita a descoberta e o acesso a documentos
- requer o estabelecimento de convenções, de modo que o acesso aos meta-dados se dê de maneira organizada e eficiente

■ Histórico

- META tags
- XML (Extensible Markup Language)
- RDF (Resource Description Framework)



EC3 - Exercício Computacional 3

- Desenvolver e implementar um agente do tipo web-robot de manutenção que
 - dado um URL correspondente a um web site determinado, vasculhe todos o links a partir deste site, e verifique se o link continua válido
 - percorra todas as sub-web pages do site, efetuando a mesma atividade
 - envie uma mensagem de e-mail para um endereço designado quando links inválidos forem encontrados