Web crawler com Banco de Dados Referencial

Carlos Roberto das Chagas Junior
Emerson de Sousa Barros
Octavio Luis Magela Oliveira
Wagner Nunes da Silva.
Centro Universitário SENAC, São Paulo, SP

RESUMO

Web crawler é um programa que navega pela WEB de forma metódica e automatizada. O processo que um Web crawler executa é chamado de web crawling ou spidering.

PALAVRAS-CHAVE: Web Crawler; Web crawling; Spidering.

Web crawler

Web crawler é um programa que navega pela WEB de forma metódica e automatizada. O processo que um web crawler executa é chamado de web crawling ou spidering.

Os web crawlers são principalmente utilizados para criar uma cópia de todas as páginas visitadas para um pós-processamento por um motor de busca que irá indexar as páginas baixadas para prover buscas mais rápidas. Crawlers também podem ser usados para tarefas de manutenção automatizadas em um Web site, como checar os links ou validar o código HTML.

Objetivo

Criar uma ferramenta de visualização de links retornando ao usuário o estado dos links e assim o mesmo poderá verificar erros no site.

Metodologia

No nosso caso o crawler começa com uma URL digitada pelo usuário (Ex: http://www.google.com), e a partir desta URL ele identifica todos os links daquela URL. Após ele ter coletado os links ele verifica coisas básicas dá pagina como tempo de resposta entre o cliente e o servidor, tempo de carregamento da página, verifica os estados dos links gerados (mensagens de erro 1XX “Mensagem informativa”, 2XX “Mensagem de sucesso”, 3XX “Mensagem de Redirecionamento”, 4XX “Mensagem de erro de cliente”, 5XX “Mensagem de outro erro. Ex: 502 Bad Gateway”).

REFERÊNCIAS

Web Crawler, http://pt.wikipedia.org/wiki/Web_crawler
Python Algorithms, Magnus Lie Hetland
Python Cookbook,2nd Edition, **Alex Martelli, Anna Martelli Ravenscroft & David Ascher**