Web crawler e utilização de ferramentas de visualização - Snooper

LEONARDO A. MENDES DOY, LUCAS T. AGOSTINHO, RODRIGO M. DA PAIXÃO.

Centro Universitário SENAC – Campus Santo Amaro

Departamento de Ciência da Computação

Av. Engenheiro Eusébio Stevaux, 823 – Santo Amaro – CEP 04696-000 São Paulo (SP).

leonardo.doy@gmail.com, teles@pwi.com, rodrigomendonca@pragmapatrimonio.com

Abstract:

We developed in this project a site, where we used the concepts about web crawlers and visualization tools. For the web crawler, we implemented the Python language and we used a HTML5, JavaScript, CSS3 and Ajax. The main objective of the project is provide a search tool where it facilitates the visualizations of links, images and archives from a determinate site.

Key Words: web crawler, Python, HTML5, CSS3, JavaScript.

Resumo:

Neste trabalho desenvolvemos um site, onde usamos o conceito de web crawler e utilizamos ferramentas de visualizações, para mostrar as buscas. Para a realização do web crawler, foi implementada a linguagem Python e no site utilizamos o HTML5, JavaScript, CSS3 e Ajax. O principal objetivo do projeto é proporcionar uma ferramenta de busca que facilite a visualização de links, imagens e arquivos de um determinado site.

Palavras Chaves: web crawler, Python, HTML5, CSS3, JavaScript.

1. Introdução

Com a evolução internet, as necessidades de acessos aos conteúdos da Web ficaram cada vez mais intensas e a procura desses mesmos conteúdos cada vez mais difíceis, pois o ambiente Web expandiu de uma forma assustadora. Para otimizar e sanar esse problema, foram criados os web crawlers.

Os webs crawlers surgiram logo após o aparecimento da internet. Sua função é buscar qualquer informação na rede, apresentando os resultados de uma forma organizada, e também com a proposta de fazer isto de uma maneira rápida e

eficiente. É iniciado com uma URL onde o crawler poder recuperar informações da Web se seguir por outras URLs encontradas na URL inicial.

Este projeto tem como objetivo apontar as características e funcionalidades de um Web Crawler junto com as ferramentas de visualizações, com o intuito de obter informações de sites, como por exemplo: links, imagens e arquivos. Assim, apresentamos o Snooper à todos.

2. Conceitos Básicos

Para entender melhor o mundo dos web crawlers, vamos ver algumas breves definições das linguagens utilizadas para a criação do projeto.

- **HTML5:** é uma linguagem de marcação utilizada para produzir páginas na Web.
- **CSS:** é uma linguagem de estilo utilizada para definir a apresentação de documentos escritos em uma linguagem de marcação, como HTML ou XML.
- **JavaScript:** principal linguagem para programação client-side em navegadores Web.
- **Python:** é uma linguagem de programação de alto nível, interpretada, imperativa, orientada a objetos, de tipagem dinâmica e forte.
- **Ajax:** é o uso metodológico de tecnologias como JavaScript e XML, providas por navegadores, para tornar páginas Web mais interativas com o usuário, utilizando-se de solicitações assíncronas de informações

3. Snooper

O Snooper foi desenvolvido a partir dos conceitos que obtivemos, através de pesquisas, sobre os web crawlers e com as ferramentas de visualizações. Ele é capaz de vasculhar um determinado site e retornar imagens, links e arquivos (se o site possuir).

Fig.1 – Logotipo do aplicativo.

4. A importância das ferramentas de visualização

Nem sempre um aplicativo bem programado, com eficiência e rapidez pode ser considerado "bom" no mercado se não possuir uma interface bonita, amigável e de fácil interação, por isso a importância de como manusear as ferramentas de visualização.

A ferramenta deve ser usada sempre para maximizar nossas habilidades, e o uso de computadores deve ser o mais simples, seguro e agradável possível. Criação de sistemas difíceis de usar pode inviabilizar o sucesso de softwares que poderiam ser bastante úteis.

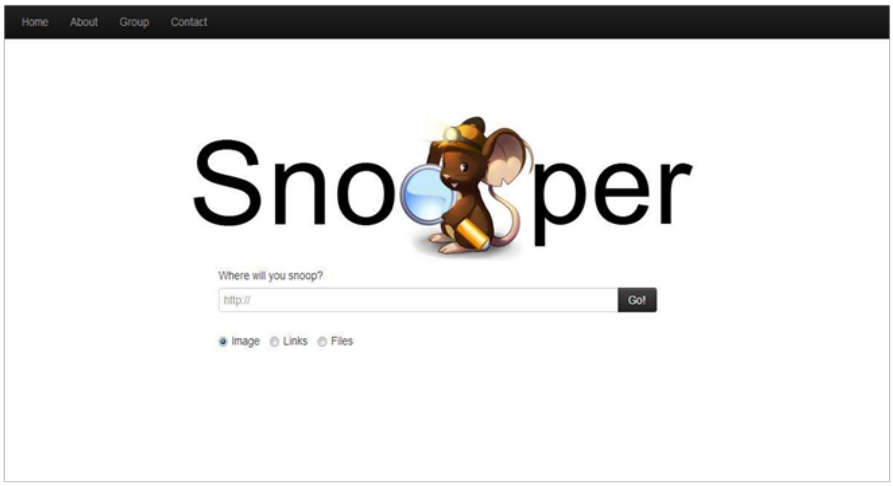

Fig.2 – Interface do aplicativo.

5. Web Crawler

Um web crawler é um programa relativamente simples e automatizado, ou script, que verifica através de páginas da Internet para criar índices dos dados que estiver procurando.

Há diversas maneiras de utilizar um crawler, mas essencialmente um rastreador da web pode ser usado por quem quer coletar informações por meio da Internet. Os motores de busca usam frequentemente crawlers para coletar informações sobre o que está disponível em páginas da web. Sua finalidade principal é coletar dados para que, quando os internautas digitarem algum termo de busca em seu site, eles possam fornecer ao usuário dados relevantes.

Quando um web crawler visita uma página web, ele consegue visualizar vários conteúdos daquela mesma página e os marca como palavra-chave. Utilizando a informação recolhida a partir do rastreador, um motor de busca, então, determina sobre quais são os índice e informações daquele site. O site é então incluído no banco de dados do mecanismo de busca e seu processo de classificação da página.

Os webs crawlers buscam qualquer informação na rede, apresentando os resultados de uma forma organizada, e também com a proposta de fazer isto de uma maneira rápida e eficiente.

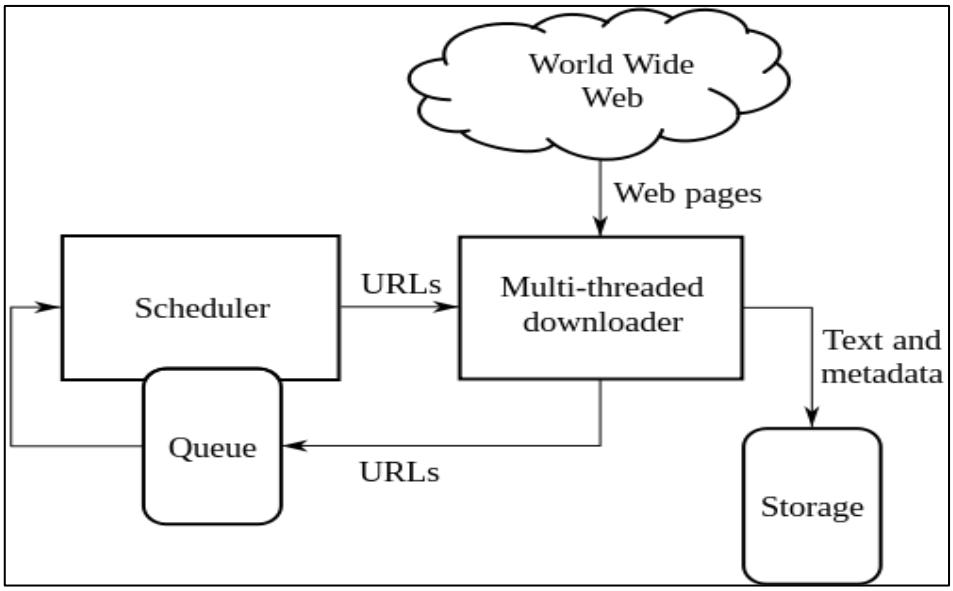

Fig. 3 – Arquitetura geral dos web Crawlers

6. Resultados

Os resultados obtidos ao efetuarmos uma busca, mostraram que o Snooper possui a capacidade de recolher, de maneira rápida e eficiente, os dados de um determinado site. De todos os sites visitados, somente aqueles que são inteiramente feitos em JavaScript o Snooper não foi capaz de trazer as informações desejadas.

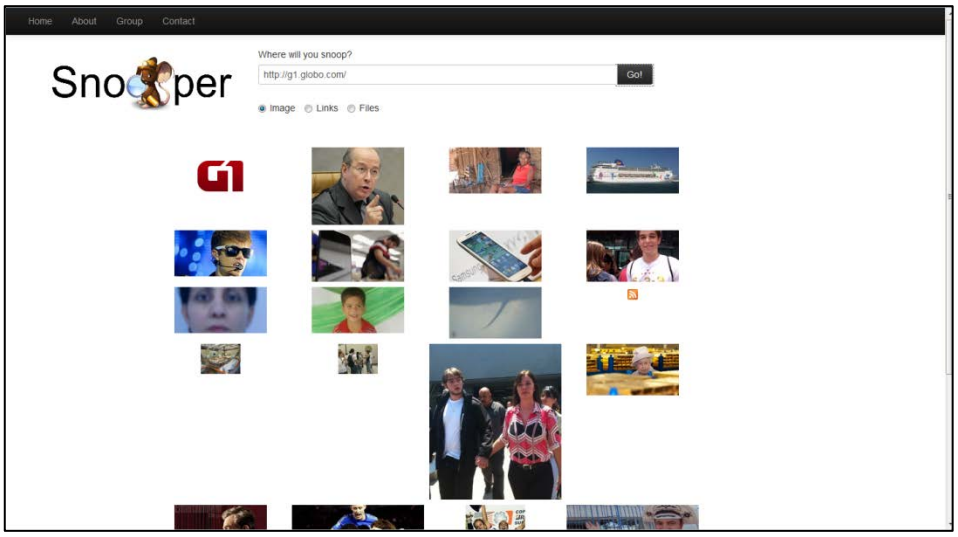

Fig. 4 – Dados obtidos pela opção "Images"

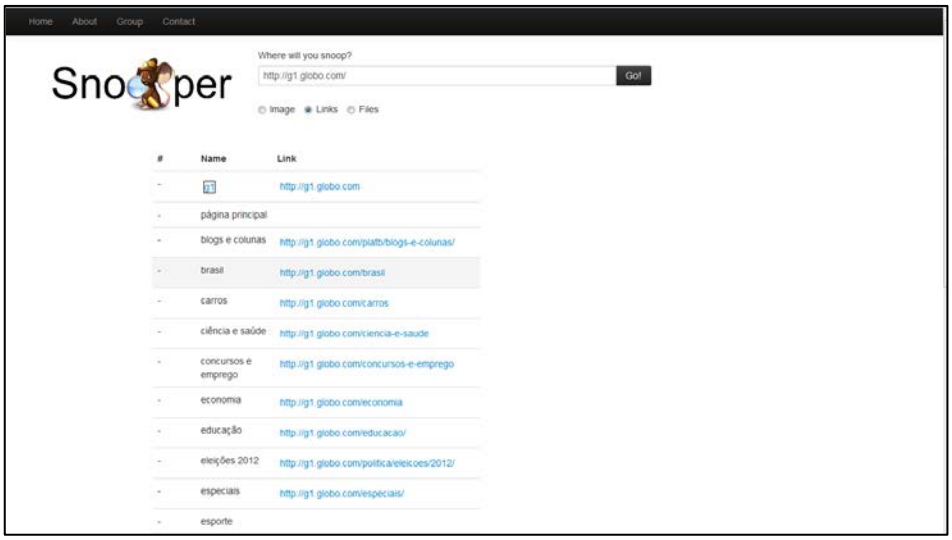

Fig. 5 – Dados obtidos pela opção "Links"

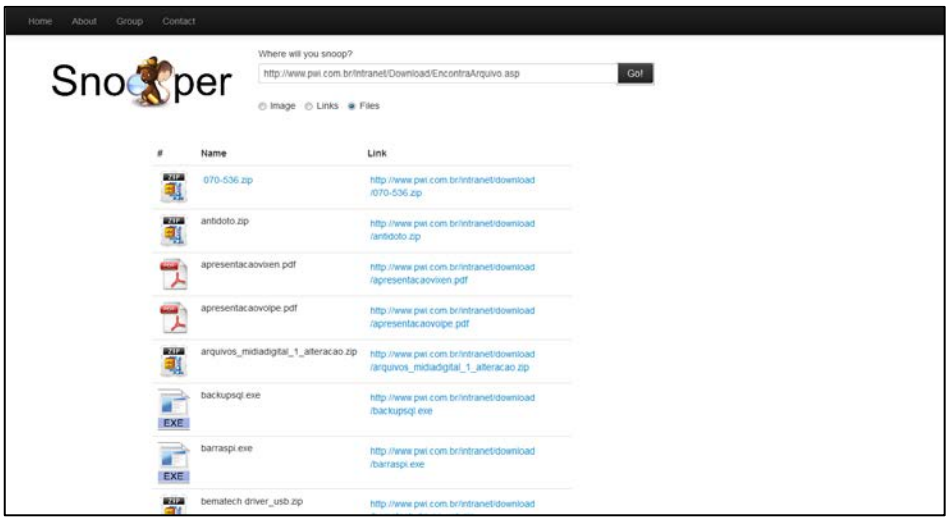

Fig. 4 – Dados obtidos pela opção "Files"

7. Conclusão

A busca de dados de um site se tornou tarefa comum e existem várias aplicações para essa funcionalidade . Neste artigo apresentamos o Snooper, um web crawler que traz links, imagens ou arquivos de um determinado site e exibe os dados de uma forma simples e intuitiva.

Com a facilidade de manuseio das ferramentas web encontrada atualmente, foi uma forma divertida e prazerosa desenvolver o Snooper. Não é necessário grande familiaridade com HTML5, CSS ou java Script, pois os conceitos são rápidos e fáceis de aprender.

8. Link Projeto GitHub

https://github.com/BCCSnooper/Snooper.git

9. Referências

- MAZANO, J. A. Estudo Dirigido Web de Javascript. São Paulo: Érica, 2001. 260 p.
- TOLEDO, S.A. Estudo Dirigido Web de HTML 4.0. São Paulo: Érica, 2001. 260p.
- Python Software Foundation. Disponível em: http://docs.python.org/3/. Acesso em: 26 de nov. de 2012
- Tuts+. Disponível em: http://learncss.tutsplus.com/. Acesso em: 10 de nov. de 2012.