Centro Universitário FEI

Modelagem de Sistemas – CC6522/CCG130

1º Semestre de 2023

Projeto 2 - Crawler

Descrição

O Web Crawler ou indexação da Web é um programa que coleta páginas da Web na Internet e as armazena em um arquivo, facilitando o acesso. Uma vez alimentado com as páginas de referência iniciais, ou "URLs iniciais", ele indexa os links da web nessas páginas. Em seguida, as páginas da web indexadas são percorridas e os links da web dentro delas são extraídos para passagem. O rastreador descobre novos links da web visitando recursivamente e indexando novos links nas páginas já indexadas.

Quais recursos estão envolvidos?

Antes de realmente construir um crawler, há algumas considerações envolvidas:

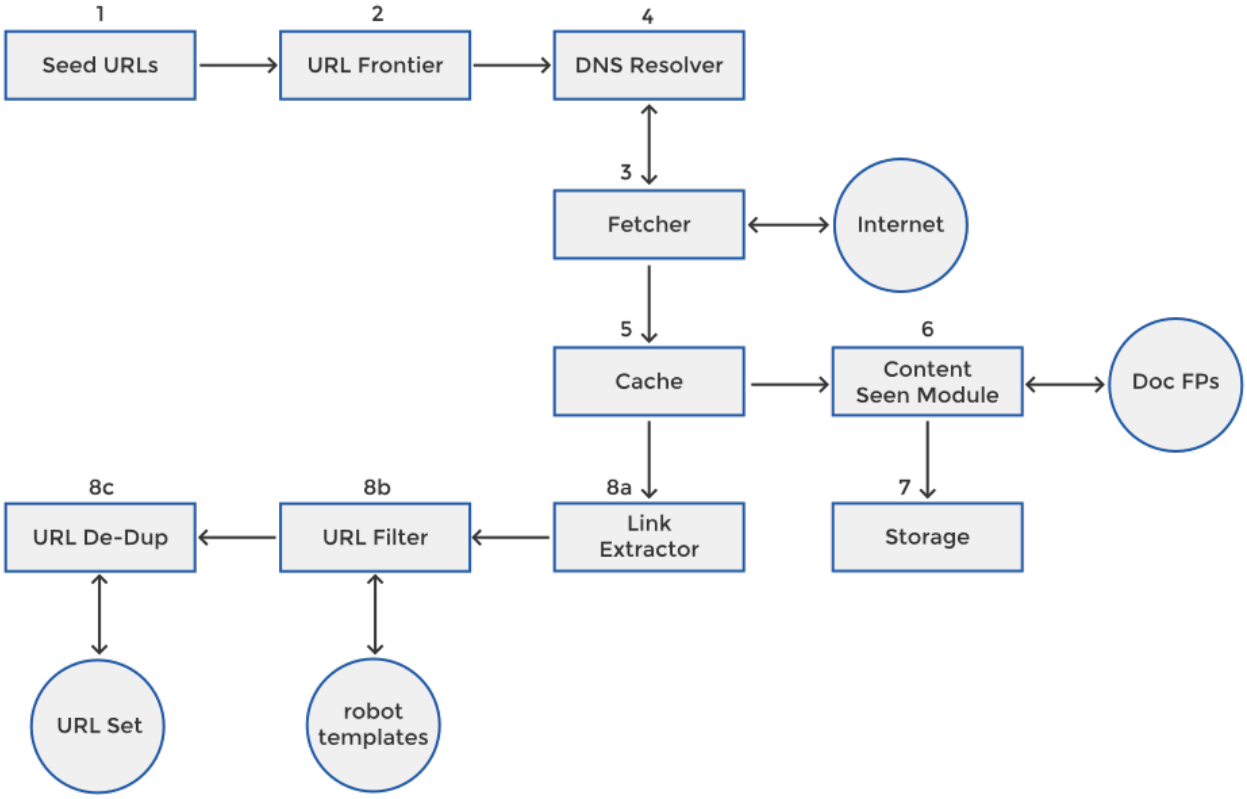
- Frequência de rastreamento: também conhecida como taxa de rastreamento ou frequência de rastreamento, refere-se à frequência com que você deseja rastrear um site. Você pode ter diferentes taxas de rastreamento para diferentes sites. Por exemplo, sites de notícias podem precisar ser rastreados com mais frequência.
- Dedup: onde vários rastreadores são usados, eles podem adicionar links duplicados ao mesmo pool de URL. A desduplicação ou detecção de duplicatas envolve o uso de um sistema com eficiência de espaço, como o Bloom Filter, para detectar links duplicados, para que seu design não rastreie os mesmos sites.
- Protocolos: pense nos protocolos que seu rastreador atenderá. Um rastreador básico pode lidar com links HTTP, mas você também pode modificar o aplicativo para funcionar em STMP ou FTP.
- Capacidade: Cada página rastreada terá vários URLs para indexar. Suponha uma estimativa de cerca de 50 bilhões de páginas. Assumindo um tamanho médio de página de 100kb: 50 B x 100 KBytes = 5 petabytes. Você precisaria de cerca de 5 petabytes de armazenamento, mais ou menos 2 petabytes, para manter as informações na web. Você pode compactar os documentos para economizar armazenamento, pois não precisará consultá-los sempre. Para determinados aplicativos, como mecanismos de pesquisa, talvez seja necessário apenas extrair as informações de metadados antes de compactá-las. Quando você precisar de todo o conteúdo da página, poderá acessá-lo por meio do arquivo em cache.

Design Diagram

Como você pode ver no diagrama de design do sistema, o loop é iniciado por meio de um conjunto de 'URLs iniciais' que é criado e inserido na fronteira do URL. A fronteira de URL aplica algoritmos para criar filas de URL com base em certas restrições, priorização e polidez, que discutiremos em detalhes mais adiante na postagem.

O módulo 3, que é o buscador de URL, recebe os URLs que estão esperando na fila um por um, recebe o endereço contra ele do resolvedor de DNS e baixa o conteúdo dessa página. O conteúdo é armazenado em cache pelo módulo 5 para facilitar o acesso aos processadores. Ele também é compactado e armazenado depois de passar pelo teste De-Dup do documento no módulo 6. Esse teste verifica se o conteúdo já foi rastreado.

As páginas em cache são processadas, passando por diferentes módulos (8a, 8b e 8c) no processo. Todos os links da página são extraídos, filtrados com base em determinados protocolos e passados pelo teste URL-Dedup para ver se vários URLs apontam para o mesmo documento e descartar repetições. O conjunto exclusivo de URLs recebidos por meio dos módulos de processamento é realimentado para a fronteira de URL para o próximo ciclo de rastreamento.



Componentes de design

1. URLs de sementes de entrada

Em primeiro lugar, seu rastreador precisará de 'URLs iniciais'. Depois de receber a entrada inicial, ele continuará extraindo e armazenando dados recursivamente. Esta lista de URLs iniciais ou URLs absolutos é alimentada na 'fronteira de URL'.

2. Fronteira de URL

O trabalho do módulo 2, a fronteira de URL, é construir e armazenar uma lista de URLs a serem baixados da internet. Para rastreadores da Web focados ou tópicos, a fronteira de URL também priorizará os URLs na fila.

3. Obtendo dados

Sempre que a fronteira de URL for solicitada para uma URL, ela enviará a próxima URL da fila de prioridade para o módulo 3, o buscador de HTML. O buscador de HTML baixa o documento no URL buscado, assim que o resolvedor de DNS fornece o endereço IP (encontre os detalhes no próximo título). O rastreador baixa o arquivo com base no protocolo de rede em que o arquivo está sendo executado. Seu rastreador também pode ter vários módulos de protocolo para baixar diferentes tipos de arquivo. O buscador, também chamado de trabalhador, invocará o módulo de protocolo apropriado para baixar a página na URL.

4. Resolvedor de DNS

Antes que o buscador de HTML possa realmente baixar o conteúdo da página, uma etapa adicional é necessária. É aqui que entra a função de um resolvedor de DNS. Um resolvedor de DNS, ou uma ferramenta de pesquisa de DNS, componente 4 no diagrama, mapeia um nome de host para seu endereço IP.

Embora a resolução de DNS possa ser solicitada do servidor, levará muito tempo para concluir a etapa, devido ao grande número de URLs a serem rastreados. Em vez disso, a melhor opção é criar um resolvedor de DNS personalizado, como você pode ver no diagrama, para complementar o design básico do rastreador. Portanto, seu resolvedor de DNS personalizado fornecerá ao buscador de HTML o endereço IP do nome do host que deve ser buscado. Uma vez que tenha o endereço IP, o fetcher baixa o conteúdo da página disponível naquele endereço.

5. Cache

Em seguida, o conteúdo baixado da Internet pelo buscador é armazenado em cache. Como os dados geralmente são armazenados depois de compactados e podem ser demorados para serem recuperados, um armazenamento de estrutura de dados de código aberto, como o Redis, pode ser usado para armazenar o documento em cache. Isso torna mais fácil para outros processadores em seu design de rastreador da web buscar os dados e relê-los sem consumir tempo desnecessário.

6. Módulo Visto de Conteúdo

Outro aspecto a considerar é se o conteúdo da URL já foi visto pelo rastreador. Às vezes, vários URLs podem ter o mesmo conteúdo. Se o documento já estiver no banco de dados do rastreador, você o descartará aqui sem enviá-lo para o armazenamento.

Usaremos o módulo 6, o módulo de conteúdo visualizado ou o teste De-Dup de documentos, para que o rastreador não baixe o mesmo documento várias vezes. Mecanismos de impressão digital, como checksum ou shingles, podem ser usados para detectar a duplicação. A soma de verificação do documento atual é comparada a todas as somas de verificação presentes em uma loja chamada 'Doc FPs' para ver se o arquivo já foi rastreado. Se a soma de verificação já existir, o documento será descartado neste ponto.

7. Armazenamento

Se o documento passou no teste de conteúdo visto no módulo anterior, ele é salvo no armazenamento persistente.

8. Dados de processamento

Você pode ter vários processadores em seu design de rastreador da Web personalizado, dependendo do que planeja fazer com o rastreador. Todo o processamento é realizado no documento armazenado em cache, e não no banco de dados armazenado, pois é mais fácil recuperá-lo. Os três processadores mais comuns que estão quase sempre presentes incluem:

8a. Extrator de links

O extrator de URL ou extrator de link pode ser considerado o processador padrão. Uma cópia da página rastreada é inserida no extrator de links do Redis ou de qualquer outro armazenamento de dados na memória. O extrator analisará o protocolo de rede e extrairá todos os links da página. Os links podem estar apontando para um local específico na mesma página, uma página diferente no mesmo site, ou um site diferente.

Um conjunto de técnicas de normalização precisará ser incorporado para tornar a lista de links mais gerenciável. Os links na lista devem seguir um formato padrão para torná-los facilmente compreensíveis pelos módulos do rastreador. Você pode:

Mapeie todos os domínios filho para o domínio principal. Por exemplo, links de mail.yahoo.com e music.yahoo.com podem ser marcados como www.yahoo.com. Se os componentes do link estiverem em maiúsculas, converta-os em minúsculas Adicione o protocolo de rede ao início do link, se estiver faltando Adicionar/remover barras invertidas no final do link

8b. Filtragem de URL
O filtro de URL recebe o conjunto de URLs padronizados e exclusivos do módulo extrator de links. Em seguida, dependendo de como você está usando o rastreador da Web, o filtro de URL filtrará os arquivos necessários e descartará o restante.
Você pode criar um filtro de URL que filtre por tipo de arquivo. Por exemplo, um rastreador da Web que rastreia apenas arquivos jpg manterá todos os links que terminam com '.jpg' e descartará o restante. Além do tipo de arquivo, você também pode filtrar os links por seu prefixo ou nome de domínio. Por exemplo, se você não deseja rastrear links da Wikipédia, pode projetar seu filtro de URL para ignorar os links que apontam para a Wikipédia.
É neste ponto que podemos implementar o protocolo de exclusão do robô. Como o buscador de URL já terá buscado um documento chamado robot.txt e mapeado as páginas fora do limite para a lista de URL, o filtro de URL descartará todos os links que o site não permite o download. Discutiremos a necessidade do protocolo de exclusão de robôs mais adiante no post.
A saída do filtro de URL são todas as URLs que queremos manter e passar para a fronteira de URL após algum processamento adicional.

8c. URL De-Dup
URL De-Dup é normalmente implementado após o módulo de filtro de URL. O fluxo de URLs que sai do filtro de URL pode não ser exclusivo. Você pode ter vários URLs no fluxo que apontam para o mesmo documento. Não queremos rastrear o mesmo documento duas vezes, então um teste De-Dup é executado em cada link filtrado antes de passá-lo adiante.
Idealmente, seu rastreador deve armazenar um banco de dados de todos os URLs rastreados — vamos chamá-lo de conjunto de URLs. Cada URL a ser testada é mapeada para cada uma das URLs no conjunto para detectar uma repetição.

Passe rapidamente por suas entrevistas de codificação com Hacking the Coding Interview.

O Filtro Bloom
O filtro Bloom é uma opção com uso eficiente de espaço para verificar se uma URL já está presente no banco de dados. Se já estiver no banco de dados rastreado, é descartado e, caso contrário, é passado para o próximo módulo. No entanto, o filtro Bloom nem sempre é confiável. Embora um falso negativo não seja possível, há chances de um falso positivo. Se o filtro Bloom decidir, erroneamente, que a URL já está presente no conjunto, a URL não será passada para a fronteira de URL. Não é um grande problema porque você pode encontrar o mesmo URL em uma das próximas iterações de rastreamento e pode ser adicionado à lista uma dessas vezes.

Repetições de Ciclo

Depois de passar pelo teste de URL De-Dup, os URLs exclusivos são salvos no conjunto de URLs e também enviados para a fronteira de URL para repetir o ciclo.