Web Crawler - Abstract

Statement of Problem

Building an automated tool/website for improving the quality of web pages. We can run it against any web page, public or requiring authentication. It has audits for performance, accessibility, progressive web apps, SEO, Website score and many more. We give the tool a URL to audit, it runs a series of audits against the page, and then it generates a report on how well the page did. From there, we can use the failing audits as indicators on how to improve the page. Each audit has a reference doc explaining why the audit is important, as well as how to fix it.

Objectives

- *Time focused crawling*: The main objective is to only crawl on a small fraction of the Web to discover the set of pages covering a certain Base URL.
- *URL Normalization*: Crawlers usually perform some type of URL Normalization in order to avoid crawling the same resource more than once. *Politeness Policy*: Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. Needless to say, if a single crawler is performing multiple requests per second Web Crawler - An Overview 267 and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers. The costs of using Web crawlers include:
    - network resources, as crawlers require considerable bandwidth and operate with a high degree of parallelism during a long period of time
    - server overload, especially if the frequency of accesses to a given server is too high
    - poorly-written crawlers, which can crash servers or routers, or which download pages they cannot handle. personal crawlers that, if deployed by too many users, can disrupt networks and Web servers**.**

- *Parallelization Policy*: A Parallel crawler is a crawler that runs multiple processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid repeated downloads of the same page.
- *Cloud Based web crawling* - Using a cloud-based web crawler means that you can login to the site from any device and any location. Not limited by users or devices, you can set a crawl going from your mobile phone and export the report on your laptop. Accessing our tool through web browsers reduces the potential for compatibility issues, runtime errors and problems with OS (Operating System) updates. Also, once you login the history of URLs crawled can be viewed along with their analysis.

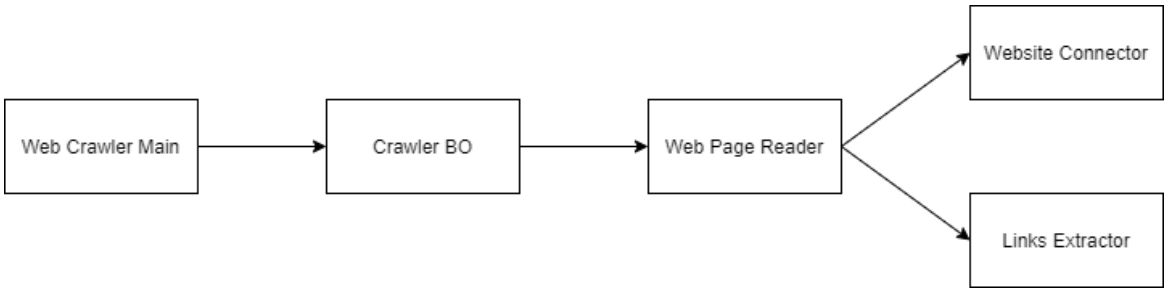
Web Crawler - Abstract

- *Info on Fixing Warnings and Errors*: Warning and errors such as HTTPs pages have internal links to HTTP, Fix Broken Links, Island Pages, href lang defined but html lang missing etc. The tool will show how critical the errors or warnings are (red - very critical, yellow - warnings/low critical, green - perfect) and also show suggestions as to how we can fix them.

Introduction to Proposed Solution

Simple workflow of the Crawler (a more detailed workflow is defined under #Architecture Diagram):

1. Request the HTML for the Page
2. Parse the page for every link for every link in the returned list, check if it's already in the crawled list
3. If it is then discard it, if not then add it to the list of links to be crawled.
4. This will continue until the number of links to be crawled is zero and thus all pages of said website have been crawled.



Novelty of Approach

Unique features:

- *Leaderboard*: We will display a leaderboard of the websites crawled and display the top 10 sites having the highest score so that other users can improve their own sites by having a look at such top websites.
- *Comparison*: We will display an option for users so that they can compare their websites with the top leaderboard site and ways to improvise it.
- *Real Time Snapshot*: Real Time snapshot of what the web page looked like at a given time.
- *Real Time Scanning*: We will display a progress bar which will show the progress of the scanning based on the number of URLs crawled out of the total URLs scanned.