Web Crawler

Saurabh Mathur `14BIT0180`
Tushar Bhatia `14BIT0163`

August 11, 2015

Abstract

A Web crawler is an Internet bot which systematically browses the World Wide Web, typically for the purpose of Web indexing. The large size and the dynamic nature of the Web makes it necessary to continually maintain Web based information retrieval systems. Crawlers facilitate this process by following hyperlinks in Web pages to automatically download new and updated Web pages. They exploit the graph structure of the Web to move from page to page.

Problem Description

Input A list of seed URLs

Output A set of visited URLs and a graph of the explored network.

Algorithm

The Web Crawler will operate as follows:

1. Fetch the first HTML page using the seed URL.
2. Parse the HTML document and add all URLs in the document to a list.
3. Fetch each URL from the list.
4. Repeat `Step 2` and `Step 3` for each of the URLs in the list.

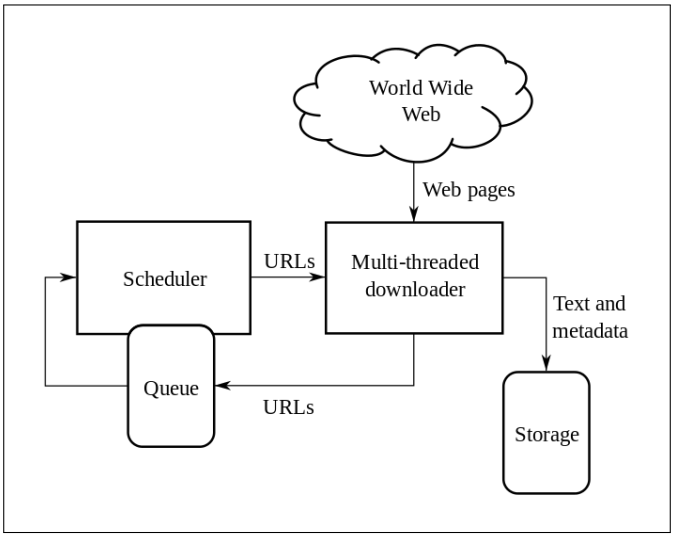

Figure 1: Architecture of the Web Crawler