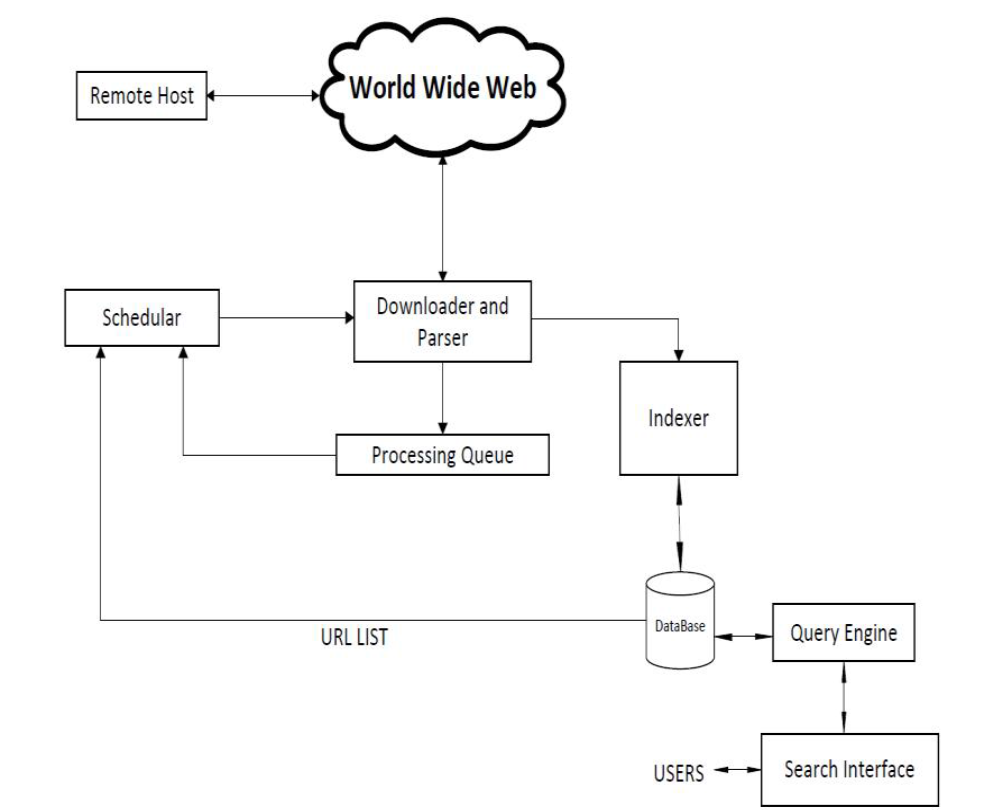
**INTRODUCTION**

World Wide Web is a collection of text documents, images, multimedia and other resources, which are linked by URLs and hyperlinks, usually accessed by web servers. According to the estimation WWW contains more than 2000 billion visible pages on web. Due to large number of pages on web, the search engine depends upon web crawlers to create and maintain indices for the web pages. A web crawler is a program which, giving one or more than one seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks present in them, and iteratively continues to download the web pages identified by these hyperlinks. Web crawlers are short software codes also called wanderers, automatic indexers, Web robots, Web spiders, ants, bots, Web scutters. In order to download a document, Crawler picks up its initial URL (seed URL) and depending on the host protocol, thus downloads the document from the Server. The search engine relies on massive collections of web pages that are acquired with the help of Crawlers, which traverses Web by following URLs and hyperlinks and storing downloaded pages in a depository that is later indexed for efficient execution of user queries

**GENERAL ARCHITECTURE OF WEB CRAWLER**

The general architecture of a crawler based search engine is shown in Fig.



Web crawlers recursively traverse and download web pages (Using GET and POST commands) for search engines to create and maintain the web indices. The need for maintaining the up-to-date pages causes a crawler to revisit the websites again and again. In general, it starts with a list of URLs to visit, called the seed URLs. As the Crawler traverses these URLs, it identifies all hyperlinks in the page and adds them to the list of URLs to be visited, called the crawl frontier. URLs from the crawl frontier are visited one by one and searching of the input pattern is done whenever text content is extracted from the page source of the web page.

The basic working of a web-crawler can be summarised as follows:

1) Select a starting seed URL or URLs

2) Add it to the Processing queue

3) Now pick the URL from the Processing queue

4) Fetch the web-page corresponding to that URL

5) Parse that web-page to find new URL links

6) Add all the newly found URLs into the Processing queue

Go to step (2) and repeat while the Processing queue is not empty

**CLASSIFICATION OF WEB CRAWLER**

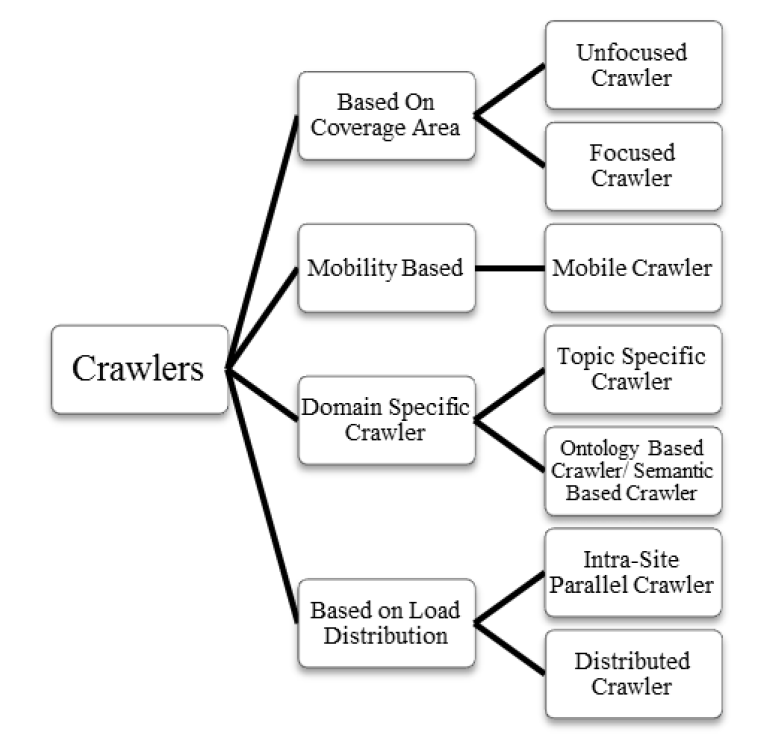
Classification of the Web Crawlers on basis of following parameters:

A. Coverage Area

B. Mobility

C. Domain Specific

D. Load Distribution



A. Based on Coverage Area

Depending upon the Web area covered by Crawlers, there are two different classes of crawlers known as (1) Unfocused and (2) Focused.

1) Unfocused Crawler

The purpose of Unfocused Crawlers is to search over the entire Web to construct their index. As a result, they deal with the laborious job of creating, refreshing, and maintaining a database of large dimensions.

2) Focused Crawler

Focused Crawler limits its function upon a semantic Web zone by selectively retrieving pages to predefined topic and avoiding irrelevant web regions to eliminate irrelevant data.

B. Mobility Based Crawler

In order to filter out the irrelevant data at the source site, where the data resides, Crawlers are transported to the site of the source. Different classes of mobility based crawler are

1) Mobile Crawler

Mobile Crawler crawl the Web using Mobile Agents. A mobile agent is an automatic-independent program act on behalf of its owner. Mobile Crawlers are transported to the remote site where they filter out any unwanted data locally before transferring it back to the search engine. These migrating crawlers remained in the remote systems and perform constant monitoring of all the web documents assigned to them for changes. Mobile Crawler reduces the network load caused by the traditional crawlers by compressing the retrieved data at remote host and reducing the amount of data transferred over the network.

C. Domain Specific Crawler

This type of crawlers traverses the web according to specific domains. Major classes of Domain Specific Crawler are:

1) Topic Specific Crawler

A Topic specific Crawler is a program used for searching information related to some specific topic from the web. The main property of topic specific Crawling is crawler does not need to collect all web pages, but selects and retrieves only relevant pages. It starts with a topic vector, and for each URL, the relevance is computed for the web page in the selected domain. If it is found to be important, it gets added to the URL list else, gets discarded.

2) Ontologies Based Crawler / Semantic Based Crawler

These crawlers make use of semantics, which helps to download only relevant pages. Semantics are provided by Ontologies. Ontologies provides a common vocabulary of an area, defines meaning of terms and relationships between them. Ontology based crawler crawl the web focusing on pages relevant to given ontologies.

D. Based on Load Distribution

In order to increase the coverage and decreases the bandwidth usage, crawlers distribute and localize the load. Depending upon load distribution two major classes of Crawler are Intra – site Parallel Crawler and Distributed Crawler.

1) Intra – site Parallel Crawler

In intra site Parallel Crawler all crawling processes run on the same local network and communicate through a high speed interconnection.

2) Distributed Crawler

Distributed Crawler is a crawler when crawling processes run geographically distant location connected by the internet.