Web Crawling

1Akash Chauhan
1Student
1Parul University

***Abstract* - The World Wide Web is interlinked collection of billions of documents formatted using HTML. Web Crawler continuously keep on crawling the web are finding any new page web pages that have been added to the web, pages have been removed from the web. Web Crawler created by Brian Pinkerton of the University of Washington and launched on April 20 1994, Web Crawler was the first Search Engine that was powered by web crawler. The crawler is a program that visit web sites and read their pages and other information in order to create entries for the search engine index. which is also known as a "Spider" or a "boot". The main purpose is Web Crawling the data is indexed by the search engine. All the main search engine, such as Google and Yahoo, use spider to build and revise their indexes. Moreover, the spider is used for automatic maintenance tasks on the web sites.**

***keywords* - bots, seeds, spider, crawler.**

I. INTRODUCTION

1.1 Introduction of Crawling

Web Crawler created by Brian Pinkerton. For the program that visit web sites and read their pages and other information in order to create entries for the search engine index. which is known as a 'Spider' or a 'boot'. Entire sites or specific pages can be selectively visited and indexed. Entries sites or specific pages can be selectively visited and indexed. Web crawlers collect information such the URL of the website, the meta tag information, the Web page content, the links in the webpage and the destinations leading from those links, the web page title and any other relevant information [1].

Example of Web Crawler

1 World Wide Web Worm
2 Google Crawler
3 Web Race

Figure-1 Web Crawler

1.2 Used of Web Crawler

A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Webin methodical, automated manner. This process is called Web crawling or spidering.

1.3 Web Crawler Work

A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. which is also known as a "spider" or a "bot". There are many types of web spiders in use, but for now, we're only interested in the bot that actually "crawls" the web. The program starts at a website and follows every hyperlink on each page So we can say that everything on the web will eventually be found and spider [2].

1.4 Software of Web Crawler

- Scrappy
- Heritrix
- HT Track
- PHP Crawler
- Apacha Nutch

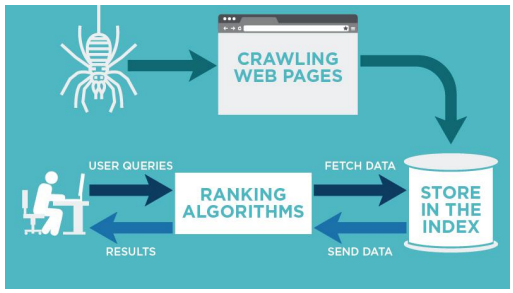

Figure – 2 - Web Crawling

1.5 Types of Crawler

- Focused Web Crawler
- Increment Crawler
- Parallel Crawler

Focused Web Crawler

Focused Crawler is the Web crawler that tries to download pages that are related to each other. It collects documents which are specific and relevant to the given topic. It is also known as a Topic Crawler because ofits way of working.

Incremental Crawler

Incremental Crawler a traditional crawler, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents.

Parallel Crawler

A parallel crawler consists of multiple crawling processes called as C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page Selection.

1.6 What are Bots and Spiders?

They are basically automated computer programs, not people, that are hitting your website. These are programs used by search engines to explore the Internet and automatically download web content available on web sites. [3].

II. Application Area

2.1 Application Development Methodology:

The web crawler application will be developed in Java 1.4.2 for many reasons. First Java's support for networking makes downloading Web pages simple. The overall tool will be developed in an incremental fashion, giving highest priority to the most important features.

2.2 Search engine use to call website crawler

Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers copy pages for processing by a search engine which indexes the downloaded pages so users can search more efficiently [4].

2.3 Types of Search Engine

- Indexing
- Calculating Relevancy
- Retrieving the Result

Indexing

Indexing is next step after crawling which is a process of identifying the words and expressions that best describe the page.

Calculating Relevancy

Search engine compares the search string in the search request with the indexed pages from the database. There are various algorithms to calculate relevancy.

Retrieving Results

The last step in search engines' activity is retrieving the results. Basically, it is simply displaying them inthe browser in an order.

2.4 Example of Crawler Based Search Engines

Most of the popular search engines are crawler base search engines and use the above technology to display search results [4].

Example of crawler base search engines.

- Google
- Bing
- Yahoo!
- Baidu
- Yandex

III. Technique of Web Crawler

3.1 Advance Crawling Technique

Web crawling is booming from being an evolving technology to become an important part of many businesses. There are many processes for this, which is a combination of different levels of crawling. The first crawlers were developed for a much smaller web, but today some of the popular sites alone have millions of pages [5].

3.2 Technique of Crawler

1. Selectively Crawler
2. Web Dynamic

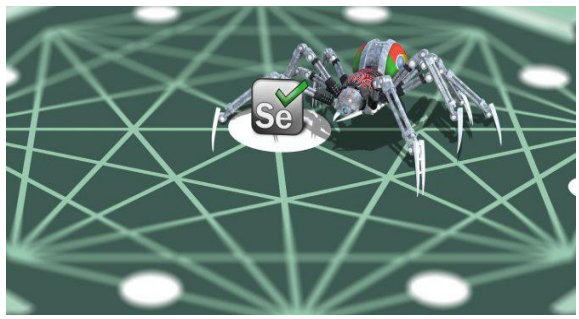

Figure - 3 Advance Crawler Technique

1 Selective Crawler

Selective Crawling is the process to retrieve web pages based on some criteria. We are using a scoring function to determine relevant contents from a page. The fetched URLs are sorted according to a relevance score from the scoring function.

2 Web Dynamic

Web Dynamics is the rate of change of information on the Web. It is mainly used by search engines for updating index.

IV. TOOLS

4.1 Top 10 Web Crawlers Tools

Web crawling (also known as web scraping) is widely applied in many areas today. It targets at fetching new or updated data from any websites and store the data for easy access. Using a web crawler tool will set free people from repetitive typing or copy-pasting, and we could expect a well-structured and all-inclusive data collection [6].

Web Crawler Tools

Octopuses
Get left
Parse Hub
Dexi.io
80Legs
HT Track
Out Wit Hub
Scraping hub
Import.io
Web Harvy

V. CURRENT LATEST WORK IN A FIELD

5.1 Top 4 Web Crawlers and Bots

1. Googlebot
2. Google app crawler (fetch resources for mobile)
3. Google AdsBot (PPC landing page quality)
4. Googlebot Video

Figure – 4 GoogleBot

1 Googlebot
Googlebot is a web crawling software search bot (also known as a spider or Web Crawler) that gathers the web page information used to supply Google search engine results pages (SERP) [7].
2 Google app Crawler
The Fetch as Google tool enables you to test how Google crawls or renders a URL on your site you can use Fetch as Google to see whether Googlebot can access a page on your site,
3 Google AdsBot (PPC landing page quality)
Google's rolling out a new system where ad landing pages will be automatically spider by a new AdsBot.
4 Googlebot Video
Googlebot-Video/1.0 is a web scraping bot used by the search engine Google that scrapes the web for videos matching users' search requests.

VI. Scope Future Enhancement

6.1 Scope Future Enhancement

Hidden Web data integration is a major challenge now days. They can neither integrate the data nor they canquery the hidden web sites. Hidden Web data needs syntactic and semantic matching to achieve fully automatic Irrigation.

1. Study of indexing technique
2. Indexing
3. User friendly search interface
4. Web 3.0 Crawling

1 Study of Indexing technique
Indexing is a data structure technique to efficiently retrieve records from the database files based on some attributes on which the indexing has been done.
2 Indexing
Indexing is the process of adding web pages into Google search. Depending upon which Meta tag you use(index or NO-index), Google will crawl and index your pages.
2 User Friendly Search Interface
As we mentioned in our last blog post about building search interfaces, having a well-designed UI is a critical component of a search installation.
3 Web 3.0 Crawling
Web 3.0 is the second phase of the Web evolution. In Web 1.0, producers created contents for the users to use it and share it. While in Web 2.0, the users equally participated in the content creation and it's sharing [8].

VII. Reference

1. https://www.techopedia.com/definition/10008/web-crawler
2. https://www.417marketing.com/how-do-web-crawlers-work
3. https://www.bounteous.com/insights/2014/08/07/bot-spider-filtering-google-analytics/?ns=l
4. https://www.webnots.com/what-are-different-types-of-search-engines
5. https://www.sayonetech.com/blog/advanced-crawling-techniques
6. https://www.octoparse.com/blog/top-20-web-crawling-tools-for-extracting-web-data
7. https://whatis.techtarget.com/definition/Googlebot
8. https://ieeexplore.ieee.org/document/5337255