International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 11
DOI: https://doi.org/10.17762/ijritcc.v9i11.5511
***Article Received**: 23 August 2021 **Revised**: 19 September 2021 **Accepted**: 24 October 2021 **Publication**: 30 November 2021*

Understanding the Concept of Different Types of Web Crawling and Its Implementation

Palika Jajoo
Assistant Professor in Computer Science,
SKIT Jaipur
palika@skit.ac.in

Abstract: - Web crawling is the method in which the topics and information is browsed in the world wide web and then it is stored in big storing device from where it can be accessed by the user as per his need. This paper will explain the use of web crawling in digital world and how does it make difference for the search engine. There are a variety of web crawling available which is explained in brief in this paper. Web crawler has many advantages over other traditional methods of searching information online. Many tools are made available which supports web crawling and makes the process easy.

Keywords: - Web Crawling, Types of Web Crawling, Advantages, Challenges, Tools.

Introduction: - Whenever a person uses internet to find out information or data specific to their interest, the search engine will respond with the results which will give all the information available about the particular request. The search engine will access all the available resources online and then provide the result. In Advance technology era, there is a process which will arrange all the related web search at one place in an organised manner which can be used by the user in future. The process of indexing the data in web page and saving it for future use is called web crawling. Spider bots, Crawler are the other words used for this technique. It will save the information using automated scripts for future use. Next time the person does similar research, then these crawlers will speed up the process and the user will get necessary information faster as compared to old search methods. This technique is commonly used for the organisations like stock market to get faster response of their search which can be related to the marketing trend and values of the stock market. With the help of web crawler this process is used to enhance the performance of the business as the business can speed up their analysis and research for any topic which will save time that can be utilised for other important tasks.

Importance of Web Crawling: - In an organisation, lot of data and information is shared and stored for future purpose. The amount of data increases day by day in an organisation or business. Similarly, huge amount of data is being produced and added online on a daily basis. In such scenarios, where the data is being added in huge volume on a daily basis, the search engine should also be such that it speeds up the search. Web crawling is one such technique which makes the searching more efficient and faster that it has many advantages over other methods. The web crawler will save the visited web pages using automated scripts in a large repository which can be further used for search in future. Next time when the user will access the same web pages it will be provided to him by web crawler large repository due to which he will get results early and faster and can finish up the tasks early. Web crawlers are also called as spider bots or simply crawlers.

Types of Web Crawler: -

1. Parallel Crawler: -This type of crawling means that at the same time various crawler can run parallel to each other to download the web pages and store them for future use. It is a convenient way of crawling as it has many advantages over single crawling doing the research.

Advantages of Parallel Web Crawling: -

- Higher download rate: - With the help of parallel crawling, the downloading and indexing of the URL becomes faster as compared to the single crawling of the web pages.

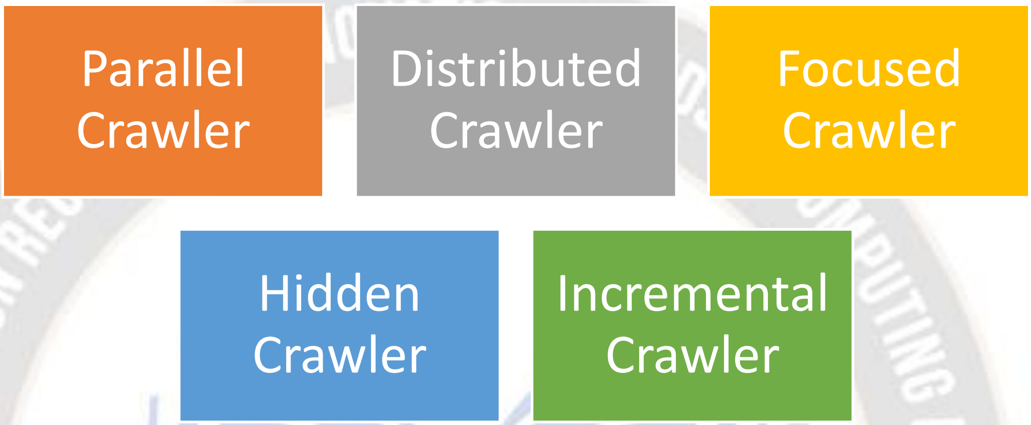

Figure 1 Types of Web Crawler.

- Reduces Load: - The parallel crawler helps to reduce the load on the network and does indexing of the URL of the web pages easily. For each type of search there will be separate crawler and can be run parallelly.
- Reliability: The web pages indexed using this king of technique is reliable and correct.

Disadvantages of Parallel Web crawling: -

- Duplicate URL: - Since there are more than one crawler searching information and indexing it, there are chances that the different crawlers have indexed same URL many times which will result in duplication of the URL.
- Occupy more space in data server: - Parallel crawling results duplicate data and URL indexing due to which the storage will need more space and most of the storage space will be wasted for indexing of same URL multiple times.
- Time Consuming process to detect duplicate indexing: - It is very time-consuming task to identify and remove the duplicate indexing from the centralised server.

2. Distributed Crawler: -In this technique, the web crawling is done through distributed devices over a network. The task of crawling is divided among the devices and systems present in the network to speed up the process of crawling to index the URL. This will help for fast indexing of the web pages as the crawling task is divided among devices available.

- Advantage: The benefit of distributing the crawling among many devices makes the crawling process easy and speed up the process. It will help to enhance the speed of the indexing of the web pages.

- Disadvantage: - The disadvantage of distributed web crawling is that if any one of the devices is down then the indexing of web page from that particular device will not be completed and hence it will result in incomplete crawling of the web pages.

3. Focused Crawler: - This type of crawling will search only specific web pages. The type of web pages being indexed using this technique will have certain specific characteristics. This type of crawling is domain specific and helps to enhance the indexing and is useful for the business where a particular study is needed. For example, companies like stock market can use this type of crawling as they need to do specific research like studying the different stock trends of various companies.

4. Hidden Crawler: - In this crawling, the spider bot will search all the available data and documents which are present deep inside the search engine. It is technique of searching the complex data which is hidden deep inside the search engine which is not easily accessed in normal simple search. The crawler will use specific algorithms and methods to search these kind of web pages for the indexing of the URL which can be used in future.

The main challenge to use this technique is that it is time consuming and the technical engineer using this technique should have the capability to use algorithms to index the complex and hidden data.

5. Incremental Web Crawling: -In this type of web crawling, the search is done in incremental manner to index the web pages and store them in centralised server for fast access in future. This means that first of all it is decided that which type of search should be done first which means that the indexing of the URL depends upon the priority of the search. Once the higher priority web pages are indexed then the crawling of other web pages are done. In this way, the crawling is done in incremental manner based upon the priority of the task.

Each of the above- mentioned web crawling technique has its own challenges and advantages. The type of the crawling technique will depend upon the type of business. The business should understand their requirement first and the according to their goals and aim should decide which crawling method to use.

Advantages of Web Crawling: - [1]

- Analysing Users Behaviour: - One of the advantages of web crawling is that it can be used by the business to understand the behaviour of the users. It can be used to track the user activities and study what are the types of data and trends which their users are following on internet. This will help the business to grow as they can provide those services which their users are searching on the internet.
- Analysing Pricing and cost of the manufacturer: - The business can use web crawling for studying the various cost and deals provided by their suppliers on their various websites. The business can have a check on all the deals easily with the help of a web crawler in order to crack the best deal provided by their supplier which otherwise would be a difficult task to do.
- Maintain directory of their contacts: - The business deals with various domains to achieve their aim a goal. There can be many departments in the business where there are number of people whose contact numbers, email ids and phone numbers should be recorded in a particular place. The business can apply crawling technique and have a list of contact details and email ids at one place for quick access.
- Fixing the prices: - Web crawling can help the business to study and compare the prices of the products provided by its competitors. After analysing the cost

offered by other companies, the business can set and fix better price options for their clients.

- Easy to find suitable candidates to fill up vacancies: - Web crawling can ease the process of hiring candidates for the vacant positions in the company. Organisation can index the URL of various job portals at one place in order to hire best candidate for their organisation. The process becomes fast and easy with the help of crawling of the job portals instead of giving advertisement for the vacancy.

Disadvantages of Web Crawling: - [2]

Besides advantages there are many challenges which a technician will handle while using web crawling for the business. Some of them are listed below: -

- Maintenance: - It is difficult to maintain the data base used in the web crawler and provide updated data to the user. This is so because many domains and people will change or modify their data once in a while. Because of this the web crawler will also need to update the web pages to provide updated version of the information. This is a big challenge which the business might face.
- The nature of data: - Internet data is vast and the type of data uploaded can be either structured or unstructured. This is the challenges in using web crawling as it will be difficult to store data which is unstructured and changes its form.
- Performance issues: - If the amount of data and search being done using web crawler is huge then the web crawler will consume large amount of network bandwidth which in turn will put unnecessary burden on the web pages being index and loaded and which in turn will result in poor performance.
- Lack of understanding of user interest: - In some cases where the crawler is not able to find proper web pages for resolving the users query then it is bound to download irrelevant web pages in response to the user’s query. This will occupy unnecessary storage in the database and will not be of much use to the user.

Tools used for Web Crawling: - [4]

- HTTrack: - With the help of this tool, website can be downloaded from the internet. It has the facility to download one or more than one website and serves many versions like windows, Sun, Linux etc.
- Cyotek Webcopy: - It is free web crawler which is used to download some part of the website or whole website for analysing it later. Before downloading it will check for any virus etc as it will download it on hard disk of the system.
- Scraper: - This tool is used for the research purpose and export the result of the research in the spreadsheet which can be used for future purposes.
- OutWit Hub: - This tool can be used to crawl small amount of data or large amount of data and can store the information in proper format.
- Visual Scrapper: - This tool is also one of the free tools available which does not require coding to crawl the web pages. It has simple UI which can be used for the data crawling from the internet.

Conclusion: - Hence it is observed that due to huge and large amount of data present in internet, web crawling helps to manage the data gathered from search engine from internet in one place. This makes the search easy and saves a lot of time for doing analysis of already existing data. There are many types of web crawling which can be chosen

DOI: https://doi.org/10.17762/ijritcc.v9i11.5511
***Article Received**: 23 August 2021 **Revised**: 19 September 2021 **Accepted**: 24 October 2021 **Publication**: 30 November 2021*

based on the type of search and data of interest of the organisation.

References: -

1.https://ninjaseo.com/benefits-of-website-crawler
2.https://understandingdata.com/the-advantages-disadvantages-of-web-scraping-data/
3.https://www.quantzig.com/blog/web-crawler-challenges-crawling/
4.https://bigdata-madesimple.com/top-20-web-crawler-tools-scrape-websites/