Experiment 4 : Website crawling OSINT tools

Aim: Utilize website crawling OSINT tools to gather a comprehensive list of URLs, internal links, and structure of the website.

1. Introduction:

WebCrawler: A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web and that is typically operated by search engines for the purpose of Web indexing (web spidering).Web crawling and web scraping are two similar concepts that can be easily confused. The main difference between the two is that while web crawling is about **finding** and **indexing** webpages, web scraping is about **extracting** the data found on one or more webpages.

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

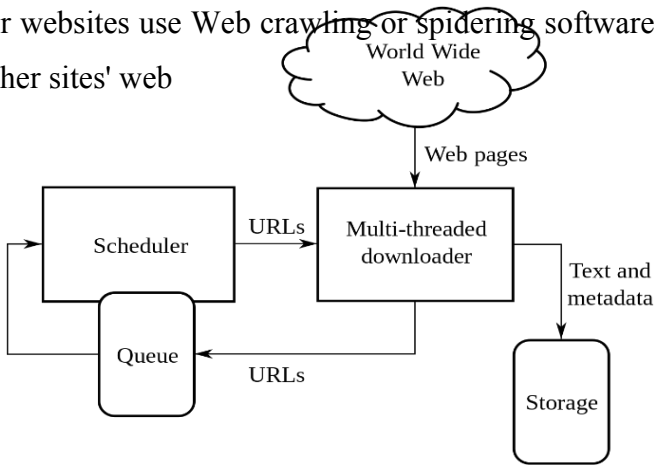


Fig 6.1: High-level architecture of a standard Web crawler

Crawlers consume resources on visited systems and often visit sites unprompted. Issues of schedule, load, and "politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For example, including a robots.txt file can request bots to index only parts of a website, or nothing at all.

The number of Internet pages is extremely large; even the largest crawlers fall short of making a complete index. For this reason, search engines struggle to give relevant search results in the early years of the World Wide Web, before 2000. Today, relevant results are given almost instantly.

Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping and data-driven programming.

2. Objective:

The few objectives of this website crawling OSINT tools:

- to gather a comprehensive list of URLs, internal links, and structure of the website.
- to learn more about an individual or a business.
- to identify information
- to learn what (almost) every webpage on the web is about, so that the information can be retrieved when it's needed.

3. Methodology:

Usually, finding information on the internet

- Begin your search with general terms and concepts. Check the sources you found and you will find more and better terms to continue your search.
- Narrow your search, you will get better results.
- Learn how Google's advanced search works.

But,

To access WebCrawler, open your preferred web browser and navigate to https://www.webcrawler.com

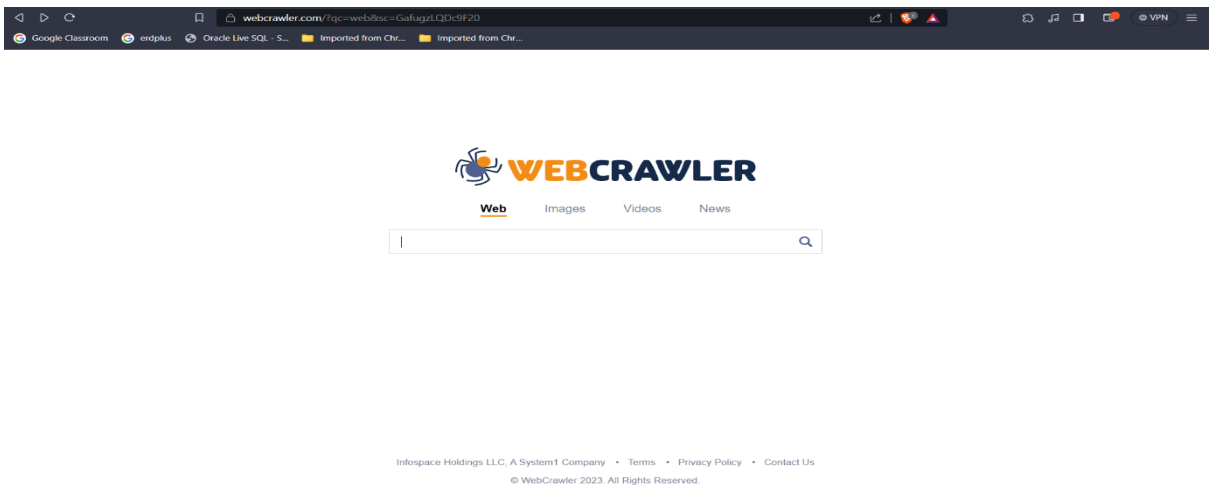


Fig 6.2: Webcrawler home page

Search for web to gather information:

WebCrawler crawls the webpages at those URLs first. As they crawl those webpages, they will find hyperlinks to other URLs, and they add those to the list of pages to crawl next. Given the vast number of webpages on the Internet that could be indexed for search, this process could go on almost indefinitely.

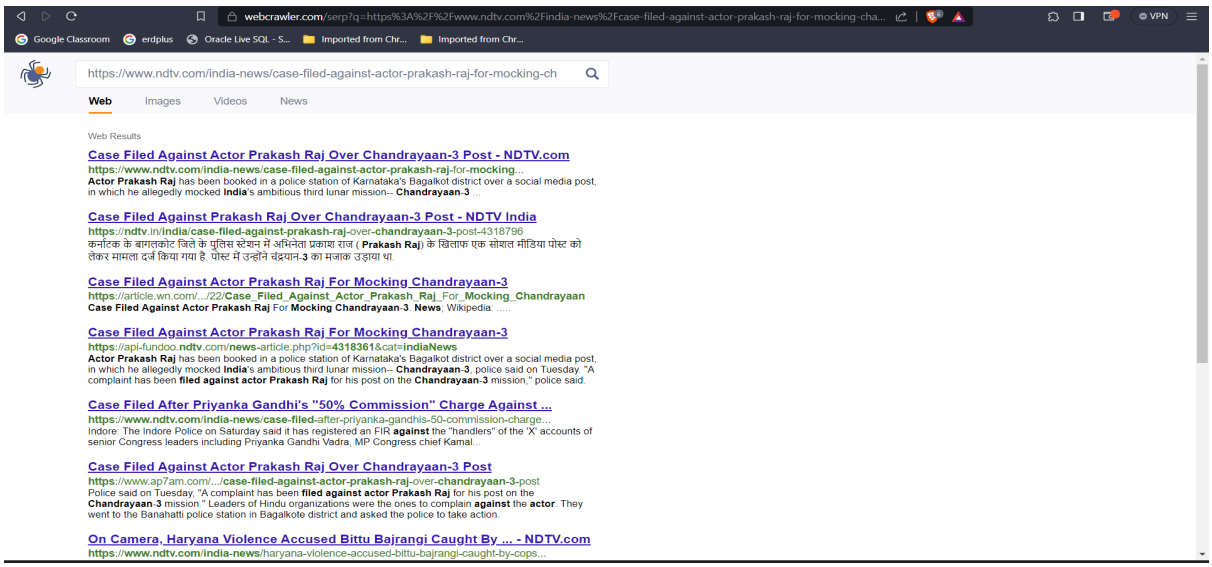


fig 6.3: Search for web to gather information

Search for media info: These too work similar to Search for web to gather information but information provided here will be media(image/video) for which query is fired.

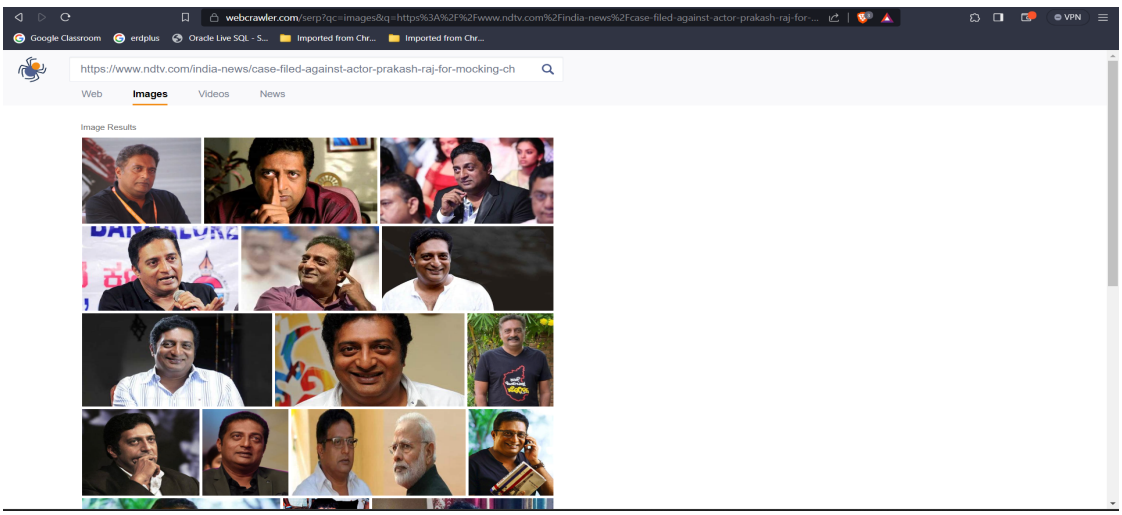


fig 6.3: Fig 6.4 Search for media information

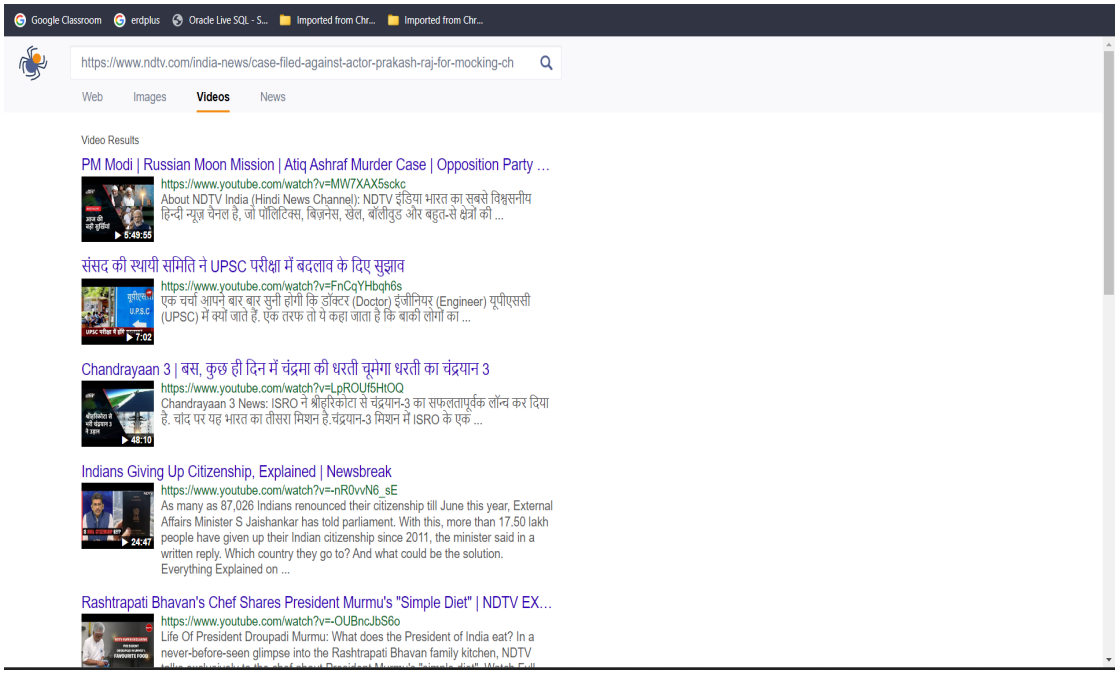


Fig 6.5: video links for web crawling

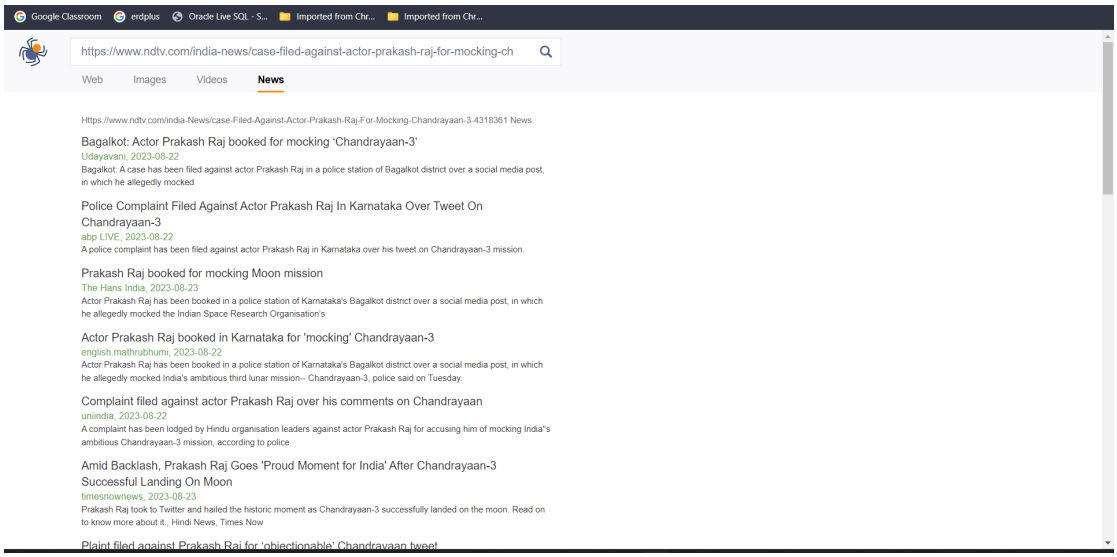


Fig 6.6: News links for web crawling

There are many tools available for web crawling. Archive it is one of free and Open Source Software, enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities

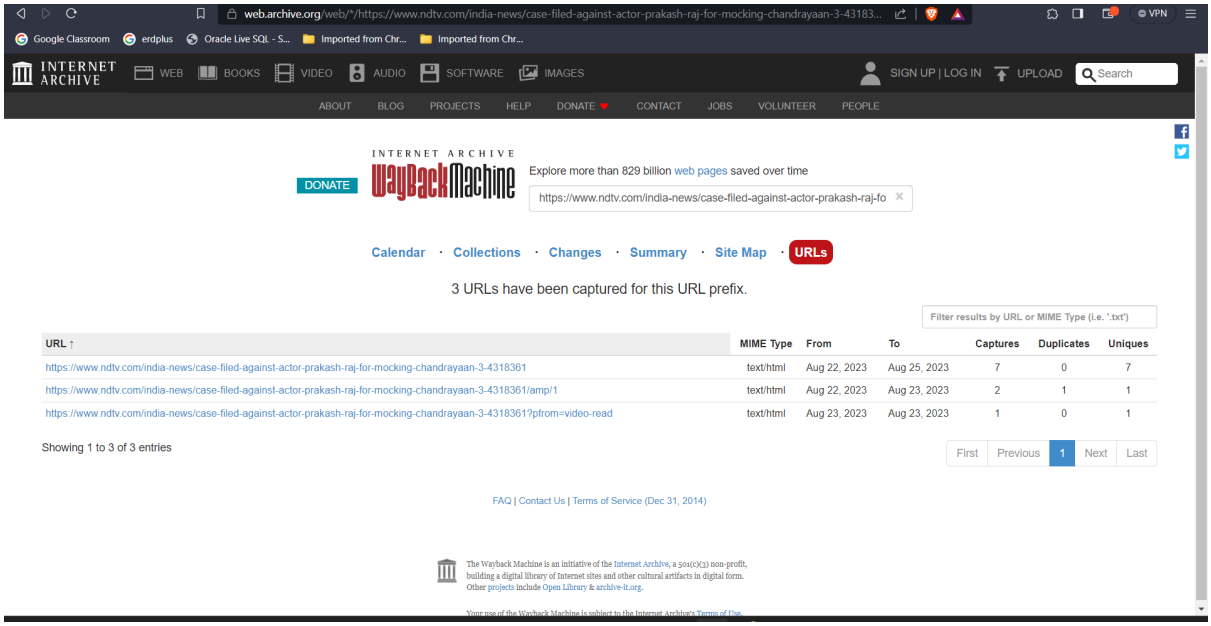
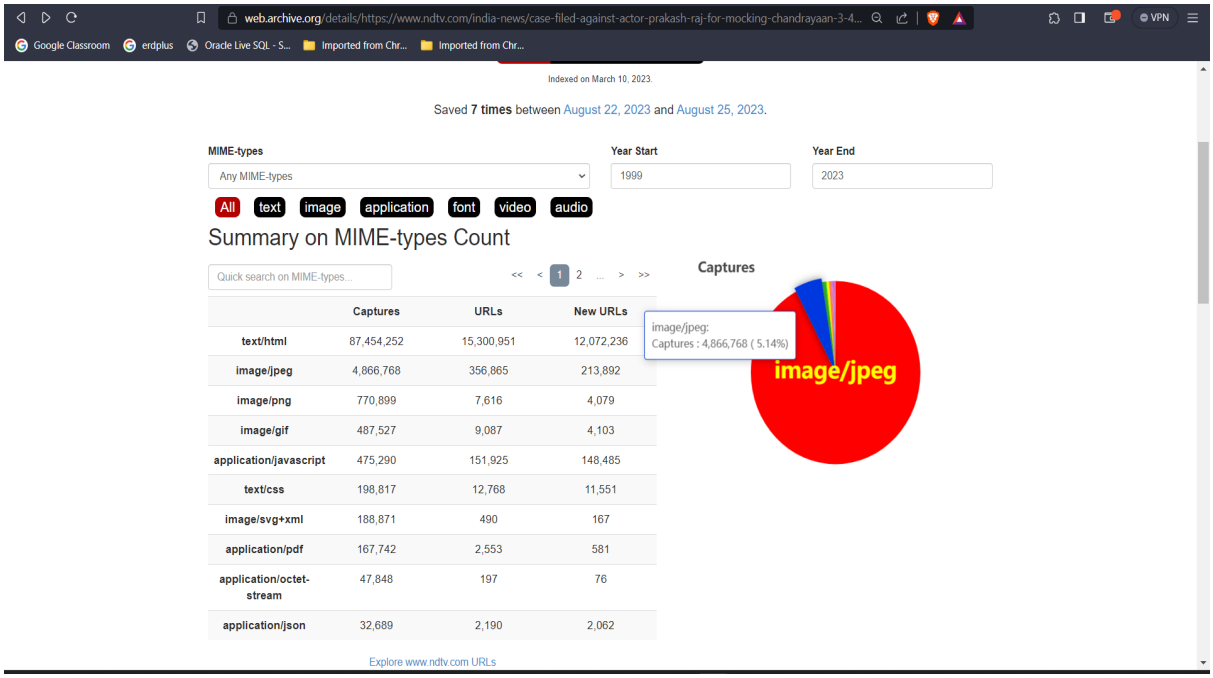


Fig 6.7: home page of Archive.it



	Captures	URLs	New URLs
text/html	87,454,252	15,300,951	12,072,236
image/jpeg	4,866,768	356,865	213,892
image/png	770,899	7,616	4,079
image/gif	487,527	9,087	4,103
application/javascript	475,290	151,925	148,485
text/css	198,817	12,768	11,551
image/svg+xml	188,871	490	167
application/pdf	167,742	2,553	581
application/octet-stream	47,848	197	76
application/json	32,689	2,190	2,062

Fig 6.8: Application of Archive.it

4. Functionalities and Benefits:

- **Comprehensive Data Gathering:** WebCrawler.com's web crawling capabilities allow it to gather data from a wide range of sources, enabling users to access a broad spectrum of information relevant to their research or analysis.
- **Time Efficiency**: The tool's automated web crawling process significantly reduces the time required to manually search for and collect data from multiple websites.
- **Data Enrichment**: By aggregating data from various sources, WebCrawler.com provides users with a comprehensive view of a topic, which can aid in making informed decisions or formulating insights.
- **Competitive Analysis:** Businesses can utilize the tool to monitor competitors' websites, track industry trends, and identify emerging opportunities.
- **Content Curation**: Content creators can leverage WebCrawler.com to curate information, gather references, and discover content ideas for their projects.
- **Market Research:** Researchers can use the tool to gather consumer feedback, opinions, and sentiments from online discussions and forums.

WebCrawler.com finds applications in various fields are as follows:

1. **Market Research:** Companies can use the tool to gather information about competitors, industry trends, and customer preferences from publicly available websites.
2. **Media Monitoring**: Media outlets can employ the tool to track mentions, sentiment, and coverage of specific topics across online news sources.
3. **Academic Research:** Researchers can use WebCrawler.com to collect data for studies related to linguistics, social sciences, and other disciplines.
4. **Cybersecurity:** The tool can be utilized to identify vulnerabilities by analyzing website content and extracting information that might expose potential weaknesses.

5. Conclusion: WebCrawler.com plays a crucial role as an OSINT tool by enabling efficient and automated web crawling for data extraction. Its intuitive interface, search capabilities, data presentation, and exporting functionalities contribute to its usefulness in various domains such as market research, competitive analysis, academic research, and content creation. By harnessing the power of WebCrawler.com, users can access and utilize vast amounts of publicly available information to make informed decisions and gain valuable insights.

6. Reference:

1. Web crawler - Wikipedia: https://en.wikipedia.org/wiki/Web_crawler

2. High-Speed Site Crawler - Clear Interface & Helpful Data: www.semrush.com/Site_Audit/Ultimate_Tool
3. What is a web crawler? | How web spiders work | Cloudflare: https://www.cloudflare.com/learning/bots/what-is-a-web-crawler

4. What Is a Web Crawler, and How Does It Work? - How-To Geek: https://www.howtogeek.com/731787/what-is-a-web-crawler-and-how-does-it-work

5. What is a Webcrawler and where is it used? – GeeksforGeeks: https://www.geeksforgeeks.org/what-is-a-webcrawler-and-where-is-it-used