Design and Implementation of Search Engine with Intelligent Web Crawler

Neetu Anand, Abhinav
Department of Computer Application
Maharaja Surajmal Institute,
Affiliated to GGSIPU, Delhi, India

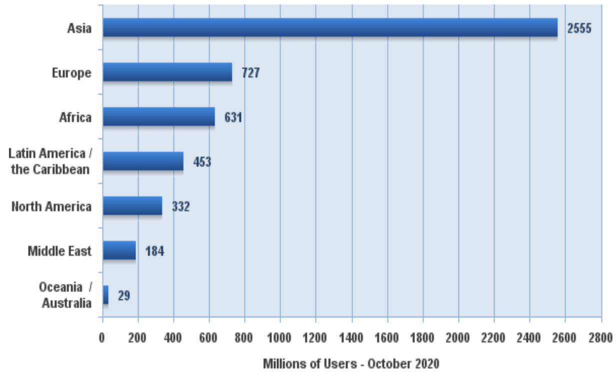
Abstract- Information Retrieval deals in searching and retrieving of information which is stored in documents and it searches the web databases and the Web. A Web crawler is a program which moves around the Internet and saves web documents in a functional way. Using these features as the base, crawler is categorized into three types of techniques: General Purpose Crawling, focused crawling and Distributed Crawling. This paper deals with the history of Web Crawlers, its use in search engines, Design and scope of development in the future with potential problems. The Research work covers two major areas: - Populating a Database (With Web Crawler), Searching from the data store. The accountability of the foremost unit is to move like spider from one location to another and to find the keywords associated with each web page. The second module is web based and focuses on the search function from the database maintained by the first module.

Keywords- Word Stemming, Web Page Crawling, Parsing techniques, Clustering of objects, Search Engines.

I. INTRODUCTION

Providing complete freedom to the server to share data that exists on the internet, the World Wide Web is extremely powerful internet-client server architecture. The organization of the knowledge is in a Hyper-Text Document, as a non-linear text system which is distributed and outsized.

Internet Users in the World
by Geographic Regions - 2020 Q3



Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 4,929,926,187 Internet users estimated in October 27, 2020
Copyright © 2020, Miniwatts Marketing Group

Fig 1. Internet User in the World. [3]

Here, pieces of texts or images are connected to other documents through anchor tags and are characterized as hypertext constituents of a document.

A standard way of retrieval and presentation of hyperlinked documents is by the employment of HTTP and HTML where servers for a required page are found with internet browsers which make use of search engines [1,2]. Ultimately, the pages that the server sends are processed on the client side.

In today's day and age, the internet is an indispensable part of a technology driven human society where a huge amount of knowledge that is present on the World Wide Web is utilised to the fullest extent. In 2020, a whopping 34.4% of the total world population, that is, 4.9 billion people are internet users, a number which rose by 1,226% from a meagre .36 billion in 2000. The proportion of Internet users within Asia itself constitutes 59.5% of the total world percentage as shown in Figure 1. Such drastic rates of growth all point towards the pivotal role that the internet already occupies, and continues to strengthen.

This rapid growth of the World Wide Web since the 1990s also estimates that the web contains about 58 billion publicly index able web documents spread across a multitude of servers. There is an obvious difficulty that one can face in terms of access to information retrieval in the vast sea of the World Wide Web which also continuously changes.

Nonetheless, there are 3 categories of tools related to Information Retrieval;

- Web Directories
- Search Engines
- Meta Search Engines

A crawler is defined as a coded script that browses the Internet in a systematic and functional manner. The Internet has a graph-like structure where access to different web pages can be done by the links demonstrated in one particular web page. In the directed graph that is the Internet, a web page is the node and the hyperlink plays the function of the edge, the whole search operation can be condensed into a process of travelling through this graph. With the help of this linked structure of the Internet, a crawler travels from the original web page to different pages.

One therefore sees how the graphical structure of the online pages is utilized by a spider to traverse from one page to another. These programs also have different names such as web spiders, worms, etc. Web crawlers therefore play the role of retrieving sites and storing them to local repositories. A reproduction of all the sites which were visited is created with the help of a spider and then a search engine [18] makes use of it, which indexes downloaded pages and help in reducing the search time dramatically. The job of the search engine is to store the information on these web pages, which are fetched with the help of the crawler.

II. LITERATURE SURVEY

The most comprehensive study of Web page change was performed by Fetterly et al [14] where estimate 151 million pages were crawled once a week for a tenure of 11 weeks, the modifications across the pages ultimately being compared. Like Ntoulas et. al. [17]; only minor changes were found wherein around 65% of all page pairs remained as is. What was also concluded in the study was that the changes in the past could be used to judge changes in the future, this page length was correlated with change. Furthermore, the top-level domain of a page was correlated with change. Researchers have been largely drawn to the aspects of changes on the Web. Two such researchers, Cho and Garcia-Molina [11] noted how such pages changed when they crawled approximately 720,000 pages once a day for 4 months. Additionally, Ntoulas et.al.[17] explored page change by downloading 154 websites once a week and collected the data for a year. Their findings pointed to how a huge number of pages remained the same based on a collection of words employed to calculate the similarities, even in pages where change was found, it was only minute. The frequency of change was not a big influence on the degree of change, but the degree of change did strongly influence the future degree of change.

About 10,000 randomized URLs and 10,000 pages from the Open Directory were crawled by Olston and Panday [12] every second for a few months. They analyzed not only the measure of frequency of change but also information longevity and only a moderate correlation was established between the two. It was with this study that fresh crawl policies which took into cognizance information longevity were introduced.

In a study consisting of changes, examined with the help of a proxy, Douglas et al. [13] observed that there was a link between the rate of re-visitation and change. However, this study did not aggressively crawl web pages for changes in different visits and was largely restricted to the internet content accessed by a limited population.

A study was conducted with its central focus on how dynamic the change was and how search engines differentiated the content for users who wanted to return to websites they had previously accessed. The first idea of a crawler that could work in parallel was proposed by Jungo Cho and Hector Garcia [11].

As we already know the web grows at an exponential speed, so it was crucial to implement a parallel crawling process, so that the download time taken by crawlers could be reduced drastically. They proposed many structures for the working of a parallel crawler and also took into account the problems which could have posed a threat to it. After taking all of these requirements, the authors then compared the architectures that were put forward using several million web pages which were saved from the web. Research in the field of crawlers has also been done by Indians, for example Ashutosh Dixit [15], had developed a model based on math which predicted how many times a crawler revisited a page. According to his model, the frequency of a crawler increased with the frequency in change of the pages up to the middle level threshold and after that the value remained constant which meant that it was not affected by the number of times a page changed but it automatically reduced itself to the lower levels once it hit the upper levels.

The crawling strategy known as Breadth-First search strategy was used by Alex Go Kwang to create and program a web spider known as Pybot. It took a URL and from that all the hyperlinks present in the web page were extracted. The recursive process is repeated until there are no new hyperlinks to be found. Pybot then uses an Excel CSV format to show the web structure of the websites it crawls. The saved pages and the structure of the web page influence.

III. HISTORY OF WEB CRAWLER

The tool Archie i.e., "Archives", which was the first Internet search Engine, came into being in 1990 and used specified public anonymous File Transfer Protocol sites to download directory listings into local files. There were subsequent creations such as that of "Gopher" in 1991 which indexed plain text documents and that of "Jughead" and "Veronica" that aided in exploring these Gopher indexes. The advent of the World Wide Web in 1990 and the usage of HTML led to various Gopher sites being transformed into properly linked websites.

The first crawler was the famous "World Wide Web Wanderer", developed in 1993[4]. Used firstly to measure the dimensions of the Web, it was subsequently also employed to find the URLs stored in the first search engine program, Wandex. "Aliweb" (Archie-Like Indexing for the Web) was also one of the initial programs where users submitted URLs of a manually constructed index.

The index contained an inventory of URLs and an inventory of user written keywords and descriptions. Initially, crawlers were involved in controversies due to their network overhead but all these issues were solved as in 1994, the Robots Exclusion protocol [5,6] was introduced which gave website controllers the power to stop spiders from fetching some parts or anything off of their websites.

In 1994, the first combined "full text" spider and search engine [9, 16] named WebCrawler was launched. It really was a stepping stone in the evolution of crawlers as it allowed users to search the content of these documents rather than limiting them to keywords. The advantage of this crawler was that it reduced the possibility of redundant results and improved the search capability of the engine.

During the same period, various commercial search engines, one being "Yahoo!" were going to be launched. It was initially a directory of websites which was maintained manually and later incorporated the accessibility of a search engine. During these early years AltaVista [8]and Yahoo! had majority of the market share. With the entry of Google in 1998, Yahoo! and AltaVista [8] lost a lot of market share as Google changed the game with its minimal and simple UI, relevant search results and kept all the spam to a minimum. Google achieved this with the help of the PageRank [10,7] software and a process known as anchor term weighting.

One more thing to be noticed was that most of the early-stage crawlers worked with small batches of data but the crawlers in use by Google, handled comparatively much more complex and large amounts of data.

IV. WORKING OF WEB CRAWLER

A crawler needs a set of URLs, termed as seed URLs. Crawlers download the pages from these URLs and within these downloaded pages itself, all the available links are stored. As these web pages are indexed in storage areas, they are retrievable as per requirement. Then a confirmation process is done to check whether the documents related to the URLs are already downloaded or not.

If they haven't been, then the crawlers process them further for downloading purposes. This is a recursive process as it is done till no more of the URLs are missing. This process leads to an average of 5-6 million pages to be downloaded by crawlers. Therefore, it is an intensive process which puts pressure on the computer too. The figure 2 given below demonstrates the internal architecture and figure 3 shows the working of a crawler:

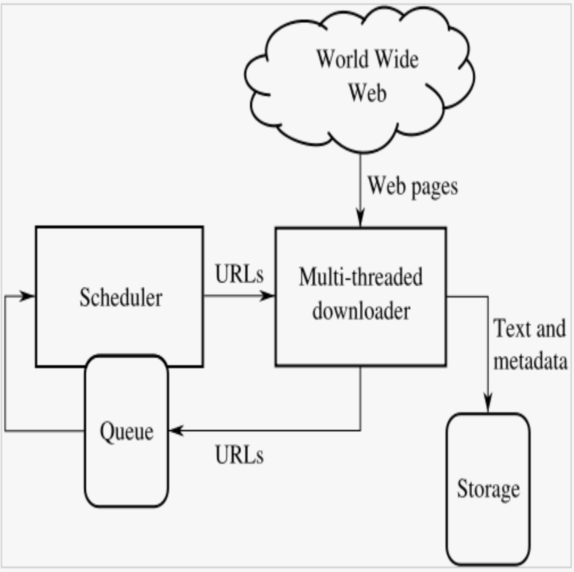


Fig 2. Architecture of Web Crawler.

The steps (shown in figure 3) that exemplify the working of a web spider are mentioned below:

- It selects a seed URL
- After selecting the URL, it is added to the frontier
- The URL is picked from the frontier queue
- Each URL fetches a web page that corresponds to it
- The crawler then searches the web page to find new links
- All the newly found links are added to the frontier
- This process is repeated until the frontier is empty

By seeing these steps, we can observe that an important function of a spider is to keep on adding new links to the frontier which is to be crawled at a later time for further processing.

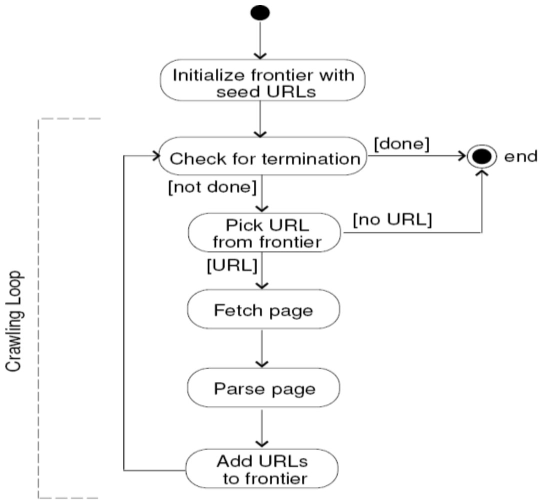


Fig 3. Steps for working of web crawler.

1. Crawling Technique:

There are three major crawling techniques: -

1.1 General Purpose Crawling: A general-purpose web crawler collects the maximum possible pages from a particular set of URLs and their links. Here, a larger number of pages from different locations are consolidated by the crawler. Owing to the nature of the multitude of pages it fetches, such a type of crawling slows down the speed and network bandwidth.

1.2 Focused Crawling: This kind of crawler functions to collect documents only on particular and specific topics, resulting in reduced amounts of traffic on a network and downloads. This crawling technique looks for pages that fit into clearly demarcated predefined requirements and crawls only the regions which are relevant leading to significant saving of hardware and network resources.

1.3 Distributed Web Crawling: It is a technique where search engines employ multiple computers for indexing the Internet with the help of web crawlers. These systems use the help of users who are voluntarily willing to offer their own bandwidth resources towards the purpose of web crawling which makes it more efficient.

2. Modern Parallel Crawler:

In today's time, search engines such as Google, Safari, Firefox etc. are not dependent on only one web crawler. Rather, they make use of multiple spiders using parallel processing to complete their work. Nonetheless, parallel connections to face problems in working such as overlapping, quality degradation and network issues. Figure 4 Illustrated how parallel crawler works:

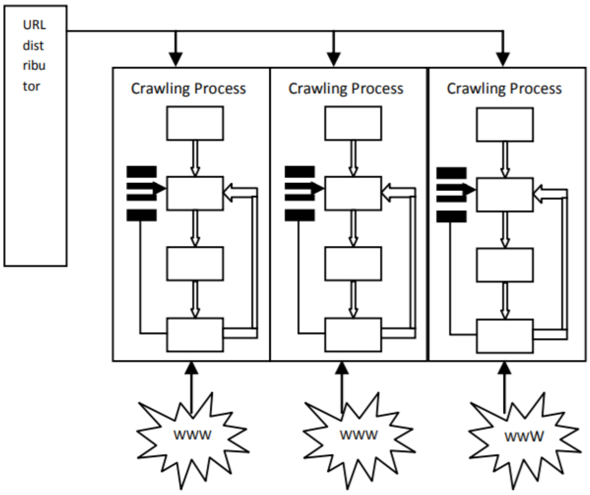


Fig 4. Working of Parallel Converter.

V. DESIGN AND DEVELOPMENT OF WEB CRAWLER

For the development and administration of Web Crawler Module, the following software and tools are required

- Visual Studio Code
- Web Browser (Internet Explorer, Mozilla Firefox, Google Chrome, etc.)
- Nodejs.
- Apache
- Puppeteer

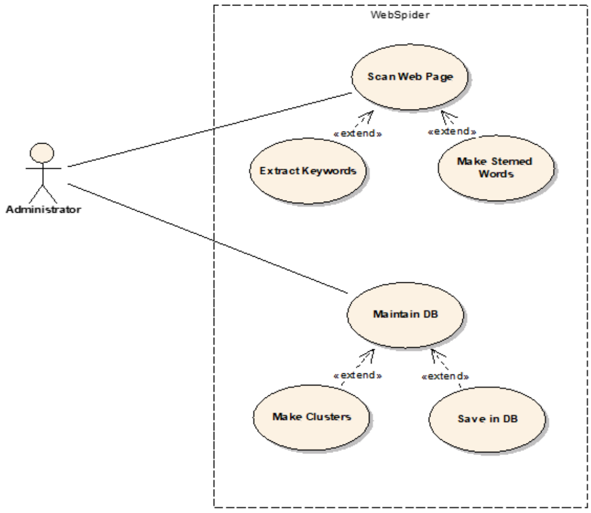


Fig 5. Use Case Diagram for Web Crawler.

The figure 5 shows the use case diagram of Web Crawler. The function of each use case is described below.

- **Use Case:** Scan Web Page: Administrator enters a URL of a web page and the system starts visiting the textual content of that web page to Extract keywords for Stemming.
- **Use Case:** Extract Key Words: All the predefined common words are dropped from the list of contents, and only the keywords remain for stemming.
- **Use Case:** Make Stemmed Word: Extracted keywords are converted to their root word.
- **Use Case:** Maintain Database: Administrator triggers an action to Save the found Stemmed words from all the URLs. The frequency of occurrence of each word at each URL will be analyzed to maintain the DB.
- **Use Case:** Make Clusters: Stemmed words and URLs are clustered on the basis of their frequency of occurrence.
- **Use Case:** Save In DB: List of clustered pairs of words and URLs are stored into record, by making a connection to the database. And if the pair is already stored into the record, it will only update the frequency of them into record.

1. Web Crawler Implementation:

The accountability of the primary segment is to move similar to a spider from one location to another and to catch the keywords associated with each web page. These keywords are collected after extrication them from a domain of words like an, that, are, etc. and then the keywords are stanched out by applying a stemming procedure. It means that programming will be converted into program, fishing to fish etc., and these words are then stored in the database. While searching from site to site, the crawler will respect the privileges set for users on the current site and it will do so by examining the robots.txt file.

The another segment is web based and focuses on the search function from the database maintained by the primary segment. It searches by taking the input from the user and before accessing the database, the input will be stemmed by applying the similar Stemming Algorithm used in the primary segment.

VI. CONCLUSION AND FUTURE SCOPE

In a fast-paced world where everything runs by the minute, the internet is considered to be such an important part of our lives because of its ability to provide information in a very limited time. This is only possible because every time a user searches for something, the relevant web pages show up without any additional spam results which make a web crawler vital in the field of web extraction techniques.

Web crawler is therefore a very important tool that travels the vast web and saves documents which are needed by users. Their main purpose is to locate these pages and save these web pages into different repositories. The main purpose of this paper was to delineate the working of web crawlers and their advancements in recent times. This article has also discussed various researches conducted by people on web crawlers.

In the future, research could be done to improve the efficiency of algorithms. The accuracy and tardiness of search engines need to be improved and the work of different crawling algorithms can be extended to increase the speed and accuracy of web crawling. One of the major issues that has to tackled is the scalability of the system and behaviour of different components. This can probably be tackled by setting up testbeds, consisting of workstations, that creates a simulation of the web using artificial web pages or a partial snapshot of the web.

REFERENCES

[1] Berners-Lee, Tim,(1996) "The World Wide Web: Past,Present and Future", MIT USA, Aug 1996, available at:http://www.w3.org/People/BernersLee/1996/ppf.html

[2] Berners-Lee, Tim, and Cailliau, CN, R(1990) "Worldwide Web: Proposal for a Hypertext Project" CERN October 1990, available at: http://www.w3.org/Proposal.html.

[3] "Internet World Stats. Worldwide internet users", available at: http://www.internetworldstats.com

[4] Maurice de Kunder, "Size of the World Wide Web", Available at: http://www.worldwidewebsize.com

[5] M. Koster. A Standard for Robot Exclusion, (1994) b. URL http://www.robotstxt.org/wc/norobots.html. http://www.robotstxt.org/wc/exclusion.html.

[6] M. Koster. ALIWEB - Archie-Like Indexing in the WEB. Computer Networks and ISDN Systems, 27(2): pp. 175–182, 1994a. ISSN 0169-7552. doi: http://dx.doi.org/10.1016/0169-7552(94)90131-7.

[7] B. Pinkerton. Finding What People Want: Experiences with the WebCrawler. In Proceedings of the Second International World Wide Web Conference, Chicago, Illinois, USA, Oct. 1994.

[8] Altavista, Mar. 2008. URL www.altavista.com

[9] D. Sullivan.(2003) Search Engine Watch: Where are they now? Search Engines we've Known & Loved, Mar. 4 2003b. URL http://searc hengine watch.com/sereport/article.php/ 21752 41

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. (1998) The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998. URL http://citeseer.ist.psu.edu/page98pagerank.html

[11] Junghoo Cho and Hector Garcia-Molina "Parallel Crawlers". Proceedings of the 11th international conference on World Wide Web WWW02", May 7–11, 2002, Honolulu, Hawaii, USA. ACM1-58113-449-5/02/0005.

[12] Olston, C. and Pandey, S. Recrawl (2008) scheduling based on information longevity. WWW '08, 437-9446, 2008.

[13] Douglis, F., A. Feldmann, B. Krishnamurthy, and J. Mogul.(1997) Rate of change and other metrics: A live study of the World Wide Web. USENIX Symposium on Internet Technologies and Systems, 1997.

[14] Fetterly, D., M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of Web pages. WWW '03, 669-678, 2003.

[15] Ashutosh Dixit and Dr. A. K. Sharma,(2010) "A Mathematical Model for Crawler Revisit Frequency", IEEE 2nd International Advance Computing Conference, pp. 316- 319, 2010.

[16] Google. Google's New GoogleScout Feature Expands Scope of Search on the Internet, Sept. 1999. URL http://www.google.com/press /pressrel/ pressrelease4.html

[17] Ntoulas, A., Cho, J., and Olston, C. (2004) What's new on the Web? The evolution of the Web from a search engine perspective. WWW '04, 1-12, 2004.

[18] Selberg, E. and Etzioni, O. (2000) On the instability of Web search engines. In Proceedings of RIAO '00, 2000.