INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 11, November 2023, pp : 162-164

# META - EXPERIENTIAL METHOD TO IMPROVE THE PRESENTATION OF WEB SYCOPHANT

**Zubair Shafiuddin[1], Lokendra Singh Songare[2]**

[1]PG Scholar, CSD, Dr. APJ Abdul Kalam University Indore, M.P, India.

[2]Assistant Professor, CSD, Dr. APJ Abdul Kalam University Indore, M.P, India.

## ABSTRACT

The internet is a tremendously big and dynamic place these days. A user uses the internet to get to specific websites. Information has been added to web pages through the use of text, videos, photos, and other processes. What a search engine is interface for retrieving data from websites. It is incredibly difficult to get to. information from a large web page archive stored in a URL. A search engine might use a web crawler and other web crawling techniques to obtain relevant sheets. A web crawler is a piece of software that lets users find and It does this by using crawling techniques to recover web content. Objective. This thesis proposes a new method for web crawlers called the single and multithreaded web crawling and indexing algorithm utilizing clustering technique. URLs found on pages that have been crawled are divided into two categories: intradomain and interdomain links. As hyperlinks rely on the kind and size of web page links that are saved in the queue URL, it initializes the weight for a particular webpage. The outcome of the experiment demonstrates that the suggested algorithm outperforms current techniques in terms of execution time. Initially, a crawling technique is used to extract the connections from a certain Uniform Resource Locator (URL), allowing the user to do hierarchical scanning for individual web links.

**Keywords.** URLs, interdomain links, web crawler, hierarchical scanning, indexing algorithm

## 1. INTRODUCTION

Big data is a vast and intricate collection of data crowds that is difficult to practice with the help of antiquated data control demos and on-hand databank organization companies. Large data sets that require enormous, diverse, and complex arrangements with challenges for loading, analyzing, and forecasting in order to promote procedures or outcomes are referred to as "big data." The process of examining vast amounts of data to reveal hidden patterns and connections is known as big data analytics. With the aid of this useful information, businesses or administrations might develop beyond the competition and gain more at ease and profound understandings.

To solve these issues Using single and multi-threaded web crawling and indexing algorithms with clustering approaches, a novel web crawling algorithm called the efficient crawling algorithm is developed. It's employed to boost the effectiveness of information retrieval. The suggested methodology is paired with page rank functionality to increase the effectiveness of web searches.

The benefit of the suggested approach is that, for scalability and resilience, it boosts time efficiency by dequeuing visited URLs from the buffer where web pages are encountered by crawlers. A dynamic hash table is used to extract duplicate URLs and increase the crawler system's dependability by preventing crashes caused by certain web pages.

## 2. OBJECTIVE OF THE WORK

The following are the proposed work's objectives:

- To suggest a single- and multi-threaded web crawling and indexing method that makes use of clustering techniques to enhance web crawlers' efficiency.
- To look into how many domains and links the suggested algorithm has crawled overall.
- To locate hyperlinks indexed by the suggested crawling technique.
- To assess the web crawling algorithm's harvesting and execution times.
- To use single and multi-threaded web crawling and indexing algorithms to increase web crawler performance.

## 3. METHODOLOGY

The proposed methodology contains following key features:

- The methodology discusses about improving the performance of web crawlers by proposing single and multi threaded crawling and indexing algorithm using clustering techniques.
- The clustering technique is applied on web pages taken from online DMOZ URL data set undergoes pre-processing where the raw data are administered to make it ready for further processing.
- Dataset contain the following topics: education, arts, computer science, health, shopping, business etc. The frontier is initialized with the seed URLs.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 11, November 2023, pp : 162-164

- The contents of pages are then examined and stored in the index. The content of the page is in HTML format or in text format.
- The single and multithreaded web crawling and indexing algorithm arranges the content according to the requirements and the result is prioritized using hierarchical clustering.
- In clustering step, grouping of the relevant pages are accomplished.
- The crawler performance is typically measured by the percentage of downloaded pages that are relevant to the topic.
- Finally obtained results of simulation work shows that our proposed algorithm gives better accuracy than the existing systems. Moreover the performance matrices of proposed algorithm such as harvest rate, harvest ratio and execution time achieved better results and compared with existing traditional crawling algorithms for showing improvement in performance of web crawlers using single and multi-threaded crawling algorithms

**WORK FLOW OF PROPOSED METHODOLOGY**

Flow-diagram of complete research methodology for entire research work is referred here in figure 1 below.

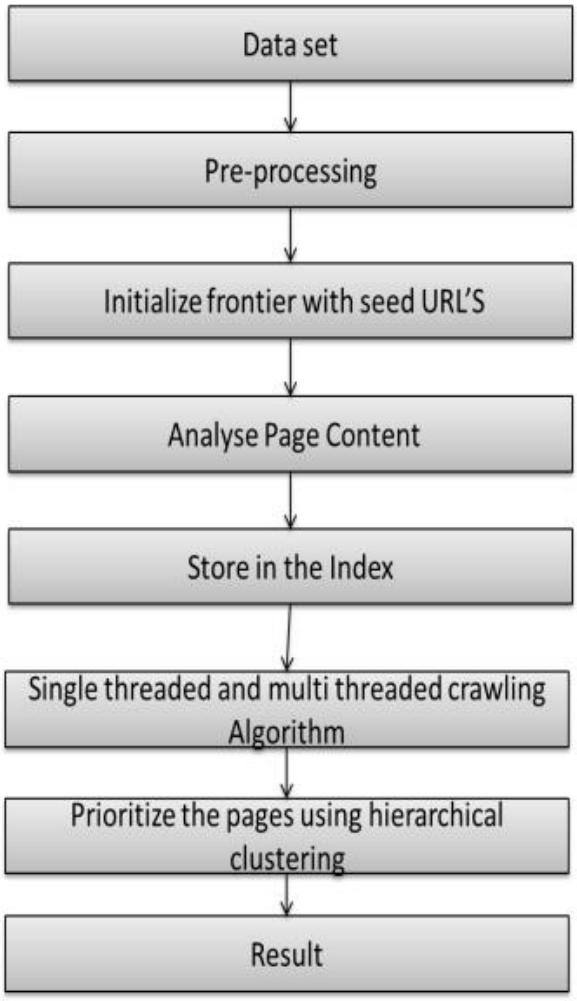


**Fig. 1:** Work Flow of Proposed Methodology

## 4. RESULT

An explanation of a minimal description length is that it is a hierarchical clustering technique that treated each input web page as a single cluster. As a result, the suggested approach evaluates the performance analysis of the clustering techniques, including Text Minimum Description Length (TXT_MDL), Min Hash Jaccard Coefficient, and Min Hash Dice Coefficient for execution and harvest time, and estimates their average execution time. When compared to the existing algorithms, the suggested algorithm yields superior results. Furthermore, for the three cluster documents, the execution times for the suggested hierarchical clustering, the coefficient TXT_MDL, the Minhash Jaccard, and the Minhash dice are 150, 300, 450, 702, 1404, 2107, 204, 408, 612, 158.5, 317, and 475.5.

As a result, the results show that the suggested approach offers more precision than the existing ones. Thus, to assess web document in website topology, the suggested method might benefit from computing the harvest ratio, harvest, and execution time.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 11, November 2023, pp : 162-164

## 5. CONCLUSION

The World Wide Web (WWW) is a network of linked content that functions as an online repository. A user uses the internet to get to specific websites. It is used to view web pages with multimedia, including text, videos, and photos, added to the content. Uniform resource locator refers to the process of browsing a web page using hyperlinks.

A separate central database system is in charge of maintaining each webpage on a search engine website. The search engine generates indexes in the web page repository in addition to the user query. A web crawler, sometimes referred to as a spider or robot, is a program that does web page crawling; a web pot is a repository where documents from web page crawls are gathered. The Crawler Frontier is its list of tasks. Its to-do list, the Crawler Frontier, is initialized with a seed URL. When new links are added to the collection of downloaded documents, the crawler accesses the webpage and eliminates them. Following the URL's removal, it determines if the user has previously downloaded any pages. The URL is reassigned to crawlers for further downloads if the document has not yet been downloaded.

Until the crawler does not leave a single URL web page for the downloading procedure, this process is repeated. Every day, web crawlers download millions of pages from the internet. Text and metadata can be stored using the scheduler, single-threaded, multi-threaded downloader, queue URL, and storage that are all included. It is responsible for maintaining and incorporating the webpage.

## 6. REFERENCES

[1] Agre, G. H., & Mahajan, N. V. (2015). Keyword focused web crawler. In 2015 2nd International Conference on Electronics and Communication Systems (ICECS) (pp. 1089-1092). IEEE.

[2] Amudha, S., and M. Phil. "Web crawler for mining web data." International Research Journal of Engineering and Technology 4.02 (2017).

[3] Bahrami, M., Singhal, M., & Zhuang, Z. (2015). A cloud-based web crawler architecture. In 2015 18th International Conference on Intelligence in Next Generation Networks (pp. 216-223). IEEE.

[4] Bailey, M. P., & Error, C. R. (2011). U.S. Patent No. 8,006,187. Washington, DC: U.S. Patent and Trademark Office.

[5] Baker, M., & Akcayol, M. (2017). Priority queue based estimation of importance of web pages for web crawlers.International Journal of Electrical and Computer Engineering,9(1), 330-342.

[6] Bhatia, K. K., & Dixit, A. (2015). Design and Implementation of Domain based Semantic Hidden Web Crawler. arXiv preprint arXiv:1509.06847.

[7] Borthakur, Dhruba. "HDFS architecture guide." Hadoop Apache Project 53.1-13 (2008):2.

[8] Cao, Fengyun, Dongming Jiang, and Jaswinder Pal Singh. Scheduling Web Crawl for Better Performance and Quality. Technical Report, TR-682-03, 2003.

[9] Castillo, Carlos, et al. "Scheduling algorithms for Web crawling." WebMedia and LA-Web, 2004.Proceedings. IEEE, 2004.

[10] Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Computer networks 31.11-16 (1999): 1623-1640.

[11] Chen, M., & Yang, X. P. (2016). Research On Model Of Network Information Extraction Based On Improved Topic-Focused Web Crawler Key Technology. Tehnicki vjesnik/Technical Gazette, 23(4).

[12] Cho, J. and Garcia-Molina, H. (2002). Parallel crawlers. In Proceedings of the Eleventh International World-Wide Web Conference. Honolulu, Hawaii.

[13] Cho, Junghoo, and Hector Garcia-Molina. "Estimating frequency of change." ACM Transactions on Internet Technology (TOIT) 3.3 (2003): 256-290.Page 116

[14] Cho, Junghoo, and Hector Garcia-Molina. "Parallel crawlers." Proceedings of the 11th international conference on World Wide Web. 2002.

[15] Cho, Junghoo, and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. Stanford, 1999.