

ADVANCED DEEPWEB CRAWLER



Dr. Ishwarya M. V¹, Dr. J. Rajeswari², S. Marlin³, Dr. Rakesh Sivalingam⁴, G. Umadevi Venkat⁵, Prinslin. L⁶, Abishek. K⁷, Akash. S⁸, Gibson. S⁹

Article History: Received: 12.12.2022

Revised: 29.01.2023

Accepted: 15.03.2023

Abstract

Web crawlers are computer programmes that browse the World Wide Web systematically, mechanically, or in an organised way. Crawling the web is an essential method for learning about and keeping up with the ever changing internet. Numerous online sites are updated often. This project presents an overview of web crawling, including onion routing and Tor network utilisation for both the surface and deep webs. This is done to monitor and preserve onion links and to keep up with the large deep web and to keep track of some illegal websites within it.

Keywords: Crawlers, TOR, Routing, Deepweb, Parallel Processing, analysis, research.

¹Head/Artificial Intelligence and Data Science Department, Agni College of Technology, Chennai

²Assistant Professor, Electronics and Communication Engineering Department, Agni College of Technology, Chennai

³Assistant Professor, Department of EEE, Agni College of Technology, Chennai

⁴Assistant Professor Grade- III, Department of Computer Applications, NITTE Institute Of Professional Education, NITTE Deemed University, Mangalore 575002

⁵Assistant Professor, CSE Department, Agni College of Technology, Chennai

⁶Assistant Professor, CSE Department, Agni College of Technology, Chennai

⁷IV year, CSE Department, Agni College of Technology, Chennai

⁸IV year, CSE Department, Agni College of Technology, Chennai

⁹IV year, CSE Department, Agni College of Technology, Chennai

Email: ¹aidshod@act.edu.in, ²rajeswari.ece@act.edu.in, ³sagayarajmarlin@gmail.com, ⁴rakesh.s@nitte.edu.in, ⁵umadevi.cse@act.edu.in, ⁶prinslinl.cse@act.edu.in, ⁷anjaanabishek10@gmail.com, ⁸leonakash6@gmail.com, gibsons572@gmail.com

DOI: 10.31838/ecb/2023.12.1.127

1. Introduction

The client-server model is used by the Internet, also referred to as the WWW. It is a robust solution that relies entirely on the server's independence when it comes to delivering web content. A hypertext document system, which is a large, dispersed, and non-linear text structure, is used to organise the data. The term "hypertextual content" in these systems refers to text or image fragments that are linked to other documents via anchor tags. HTTP and HTML offer a standardised way of retrieving and displaying hyperlinked documents. Web browsers use search engines to comb servers for the necessary information pages. The pages that the servers send are processed on the client side.

The importance of using the Internet has increased in today's society.

2. Methodology

An initial set of URLs, referred to as seed URLs, can be used by a Web crawler to begin operations. Along with the web pages for the seed URLs, they download additional new links that are present in the pages you downloaded. In order to be later recovered with the aid of these indexes as and when necessary, the pages that were retrieved are preserved and thoroughly indexed on the storage space. The extracted URLs from the downloaded page are inspected to see if the associated documents have already been downloaded. If the URLs are not downloaded, they are once more made available to web crawlers for downloading. Continue doing this until no more download URLs are required.

2.1 Existing:

Currently, the internet plays a huge role in our daily lives. The user uses the internet to do a search based on his needs. Due to the abundance and dynamic nature of web resources on the internet, producing better results that are pertinent to the search term and customising the search are the difficult problems in information retrieval. The only part of the web that is currently crawled by crawlers in this planet is the surface web. With the use of seed URLs, it will crawl the surface web and examine the URLs contained in the supplied URL. After crawling the provided URLs, it will save and index every crawled URL so that the user can use the known URLs and All of the crawled URLs from the provided URLs are indexed, allowing the user to use the well-known URLs and just conduct

a keyword search while the search engine handles the rest. A search engine's operation is intricate. A lot of content that has been scraped from numerous web pages with a tonne of unique words is indexed by search engines. Every second, they react to several inquiries. Web search engines are extremely important on a large scale, but their underlying mechanisms have received less attention.

Limitations In The Existing System:

- Can not crawl the hidden URLs.
- Reveals the IP address of the user.
- Can not crawl contact information.
- Support only some internet protocols.
- Used only on the surface web which is only a minor part.

2.2 Proposed

An alternative crawler to the current system is called the Advanced Deep Web Crawler (ADC). The deep web and black web that are present within the deep web will also be explored in addition to the surface web. In other words, the entire internet is fully crawled. With the aid of seed URLs, which might be onion URLs, surface URLs, or hidden URLs, it will crawl the web and examine the URLs to determine if any other URLs are there. Then, based on the depth value provided for the depth scan, the crawled URLs from the seed URLs are crawled once again. The crawler searches the seed URLs not only for the URLs existing there, but also for any contact information. All of the crawled URLs, seed URLs, and contact information are stored locally and indexed after the crawling process is complete and can be utilised later to access the webpage. By using a parallel processing mechanism for the crawling, the user can multitask by providing two or more URLs as the seed URL.

Above importantly, by establishing a connection with the onion-routing TOR network, the crawler offers security and anonymity.

Pro's of the proposed system:

- Easy to use.
- Provides anonymity.
- Crawls the hidden URLs.
- Support all internet protocols.
- Crawl's contact information.
- Uses onion routing on the TOR network.
- Used on both surface web and deep web.
- Supports Multitasking
- Uses Parallel processing

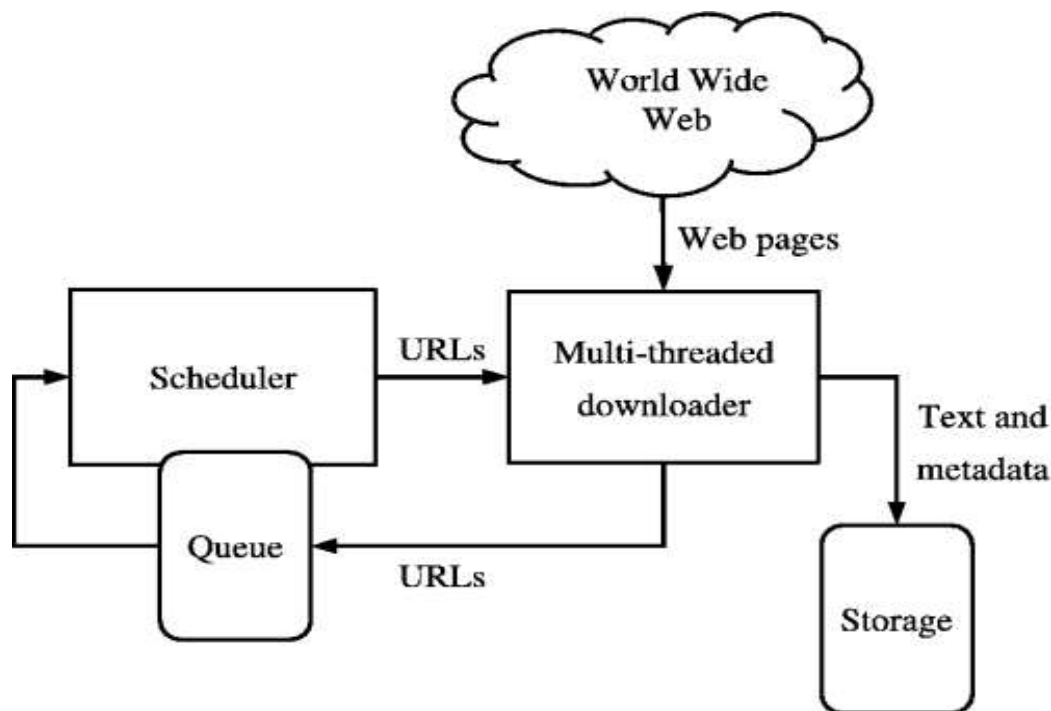


Figure 1: Working of Advanced Deepweb Crawler

3. Modeling and Analysis

The below flow diagram shows the full procedural working of the Advanced Deepweb Crawler.

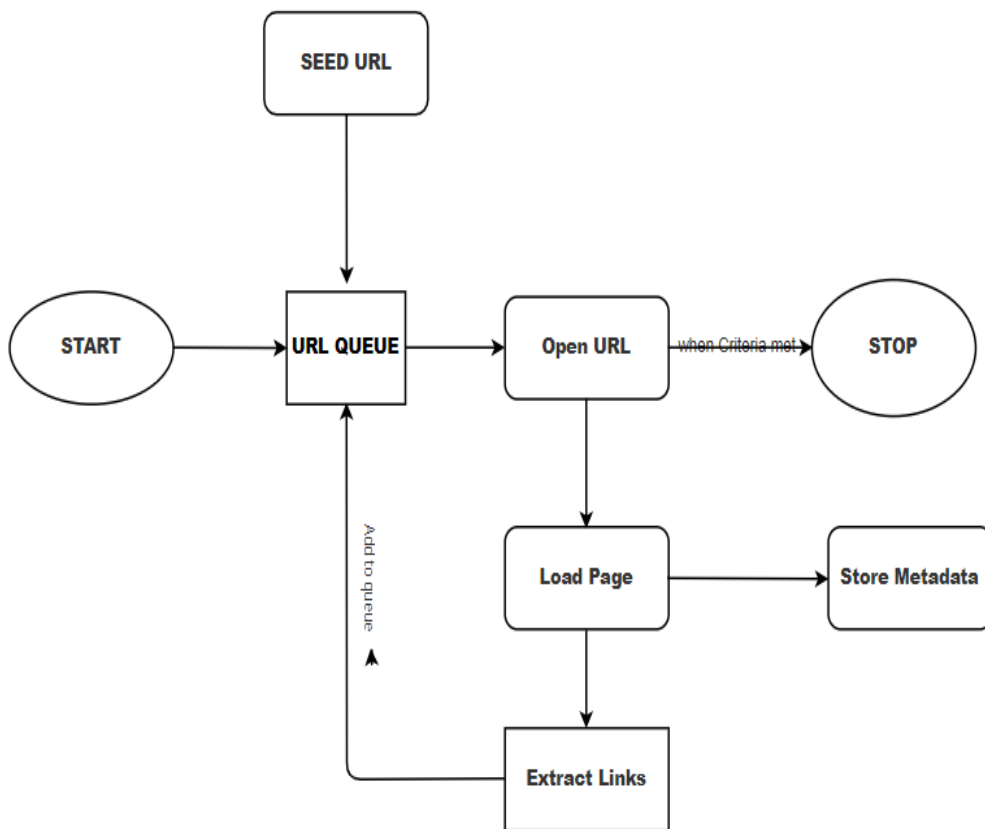



Figure 3.2: Activity Diagram

The activity diagram up above depicts the steps of the crawling process that were conducted for a single seed URL. For every distinct URL, the process is carried out separately each and every time.


```


Advanced Deepweb Crawler

From
K.Abishek & S.Gibson & S.Akash
Agni College of Technology
Under Guidance of
Dr.M.V.Ishwarya
Agni College of Technology

!!!TOR IS RUNNING!!!

!!!RUNNING CRAWLER!!!

WEBSITE: https://es-la.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://pt-br.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://fr-fr.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://de-de.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://it-it.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://ar-ar.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://hi-la.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://zh-cn.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://ja-jp.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://messenger.com/
WEBSITE: https://pay.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://www.oculus.com/
WEBSITE: https://portal.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/
WEBSITE: https://l.facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion/l.php?u=https%3A%2F%2Fwww.instagram.com%2F6h
+AT245Vp23eZ2_w6YY-Ev1231c3ny7hpC3K12Vjv08kxWq5rwnq40rxz2t9bqf0907_V6_4s5Zb676Lcsg_HYwucv5d0ub6nrYI1BanfHg_FlnyLvgU3X1FH27Mn0
a0_Kzrech8sh
WEBSITE: https://www.bulletin.com/

```

Figure 4.3:Crawling mode

```


Advanced Deepweb Crawler

From
K.Abishek & S.Gibson & S.Akash
Agni College of Technology
Under Guidance of
Dr.M.V.Ishwarya
Agni College of Technology

!!!TOR IS RUNNING!!!

<!DOCTYPE html>\n<html lang="en" id="facebook" class="no_js">\n<head><meta charset="utf-8" /><meta name="referrer" content="defau
lt" id="meta_referrer" /><script nonce="Z9HkTRMU">window._cstart+=new Date();</script><script nonce="Z9HkTRMU">function envFlush(a){f
unction b(b){for(var c in a)b[c]=a[c]}window.requireLazy?window.requireLazy(["Env"],b):(window.Env=window.Env||{}).b(window.Env)}envF
lush({"hostnameRewriterConfig":{"site":"onion","inboundName":"facebookkwhpilenexj7asaniu7vnjjbiltxjqhye3nhbshg7ks5tfyd.onion","cdnDom
ainName":"facebookcooa4ldbat4g7iactwl3p2zrf5nyuylvnhxn6kqolvojixwid.onion"},"ajaxpipe_token":"AKhXplozMEAsURkFqMI","timeslice_heartbea
t_config":{"pollIntervalMs":33,"idleGapThresholdMs":60,"ignoredTimesliceNames":{"requestAnimationFrame":true,"Event listenHandler mou
semove":true,"Event listenHandler mouseover":true,"Event listenHandler mouseout":true,"Event listenHandler scroll":true},"isHeartbeat
Enabled":true,"isArtilleryOn":false},"shouldLogCounters":true,"timeslice_categories":{"react_render":true,"reflow":true},"sample_cont
inuation_stacktraces":true,"dom_mutation_flag":true,"gk_instrument_object_url":true,"gk_log_promise_done":true,"stack_trace_limit":30
,"timesliceBufferSize":5000,"show_invariant_decoder":false,"compat_iframe_token":"AQ5G8PXu-wqxIRfKcT8","isCQuick":false});</script><s
cript nonce="Z9HkTRMU">(function(a){function b(b){if(!window.openDatabase)return;b.I_AM_INCOGNITO_AND_I_REALLY_NEED_WEBSQL=function(a
,b,c,d){return window.openDatabase(a,b,c,d)};window.openDatabase=function(){throw new Error({})}b(a)}(this)}</script><style nonce="Z9
HkTRMU"><_DEV_=0;CavalryLogger=window.CavalryLogger||function(a){this.lid=a,this.transition=!1,this
.metric_collected=!1,this.is_detailed_profiler=!1,this.instrumentation_started=!1,this.pagelet_metrics={},this.events={},this.ongoing
_watch={},this.values={t_cstart:window._cstart,this.piggy_values={},this.bootloader_metrics={},this.resource_to_pagelet_mapping={},t
his.initializeInstrumentation&&this.initializeInstrumentation()},CavalryLogger.prototype.setIsDetailedProfiler=function(a){this.is_de
tailed_profiler=a;return this},CavalryLogger.prototype.setTTIEvent=function(a){this.tti_event=a;return this},CavalryLogger.prototype
.setValue=function(a,b,c,d){d=d?this.piggy_values:this.values;(typeof d[a]==="undefined"||c)&&(d[a]=b);return this},CavalryLogger.prot

```

Figure 4.4:Extract Mode



Figure 4.7: Quiet Mode

5. Conclusion

The Internet and intranets have made a wealth of information available. The majority of individuals have access to search engines and can utilise them to locate the information they require. Because they scan the Web and retrieve web resources based on user needs, web crawlers are crucial tools for information retrieval. In order to retrieve webpages and add them to nearby repositories, web crawlers were developed. In a nutshell, crawlers are used to copy each site that is viewed. Search engines then analyse these copies, indexing the downloaded pages to enable quick searches. The review paper's main objective is to provide some clarification on past web crawling studies. This study also covers many web crawler-related studies. This results in a crawling speed of more than a million pages per day, which is adequate for the bulk of academic research projects, as we can see. While additional research is required, we predict that for larger configurations, between and pages per second and per node would be feasible by appropriately distributing components. We have provided details on our crawling system's architecture and implementation, as well as a few preliminary trials. Undoubtedly, there are a lot of methods to make the system better. Future research must focus on a crucial area: a thorough analysis of the system's scalability and the behaviour of its constituent parts. The most efficient method to do this would probably be to set up a simulation testbed

composed of several workstations that mimics the web using either artificially generated pages or a partial snapshot of the web that has been stored. We are now considering testbeds for other high-performance networked systems in addition to this. The system is being used by a number of students, who are collecting information in various ways. The primary goal of our research team is to use the crawler to examine new challenges in web search technology.

6. References

Berners-Lee, Tim, "The World Wide Web: Past, Present and Future", MIT USA, Aug 1996, available at: <http://www.w3.org/People/Berners-Lee/1996/ppf.html>.

Berners-Lee, Tim, and Cailliau, CN, R., "World Wide Web: Proposal for a Hypertext Project" CERN October 1990, available at: <http://www.w3.org/Proposal.html>.

"Internet World Stats. Worldwide internet users", available at: <http://www.internetworldstats.com>.

Maurice de Kunder, "Size of the World Wide Web", Available at: <http://www.worldwidewebsite.com>.

P. J. Deutsch. Original Archie Announcement, 1990. URL <http://groups.google.com/group/comp.archives/msg/a77343f9175b24c3?output=gplain>.

A. Emtage and P. Deutsch. Archie: An Electronic

- Directory Service for the Internet. In proceedings of the Winter 1992 USENIX Conference, pp. 93–110, San Francisco, California, USA, 1991.
- G. S. Machovec. Veronica: A Gopher Navigational Tool on the Internet. *Information Intelligence, Online Libraries, and Microcomputers*, 11(10): pp. 1–4, Oct. 1 1993. ISSN 0737-7770.
- R. Jones. Jughead: Jonzy's Universal Gopher Hierarchy Excavation And Display. unpublished, Apr. 1993.
- J. Harris. Mining the Internet: Networked Information Location Tools: Gophers, Veronica, Archie, and Jughead. *Computing Teacher*, 21(1):pp. 16–19, Aug. 1 1993. ISSN 0278-9175.
- H. Hahn and R. Stout. The Gopher, Veronica, and Jughead. In *The Internet Complete Reference*, pp. 429–457. Osborne McGraw-Hill, 1994.
- T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollermann. World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, 1(2): pp. 74–82, 1992. URL <http://citeseer.ist.psu.edu/bernerslee92worldwide.html>
- T. Berners-Lee. W3C, Mar. 2008. URL <http://www.w3.org/>
- M. K. Gray. World Wide Web Wanderer, 1996b. URL <http://www.mit.edu/people/mkgray/net/>.
- W. Sonnenreich and T. Macinta. *Web Developer.com, Guide to Search Engines*. John Wiley & Sons, New York, New York, USA, 1998.
- 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](http://dx.doi.org/10.1016/0169-7552(94)90131-7).
- M. Koster. A Standard for Robot Exclusion, 1994b. URL <http://www.robotstxt.org/wc/norobots.html>. <http://www.robotstxt.org/wc/exclusion.html>.
- 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.
- Infoseek, Mar. 2008. URL www.infoseek.co.jp
- Lycos, Mar. 2008. URL <http://www.lycos.com>
- Altavista, Mar. 2008. URL www.altavista.com
- Excite, Mar. 2008. URL www.excite.com
- Dogpile, Mar. 2008. URL www.dogpile.com
- Inktomi, Mar. 2008. URL www.inktomi.com
- Ask.com, Mar. 2008. URL <http://ask.com/>.
- Northern Light, Mar. 2008. URL <http://www.northernlight.com>
- D. Sullivan. Search Engine Watch: Where are they now? *Search Engines we've Known & Loved*, Mar. 4 2003b. URL <http://searchenginewatch.com/sereport/article.php/2175241>.
- Google. Google's New GoogleScout Feature Expands Scope of Search on the Internet, Sept. 1999. URL <http://www.google.com/press/pressrel/pressrelease4.html>
- L. Page, S. Brin, R. Motwani, and T. Winograd. 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>
- S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In P. H. Enslow Jr. and A. Ellis, editors, *WWW7: Proceedings of the Seventh International Conference on World Wide Web*, pp. 107–117, Brisbane, Australia, Apr. 14–18 1998. Elsevier Science Publishers B. V., Amsterdam, The Netherlands.
- Junghoo Cho and Hector Garcia-Molina "Parallel Crawlers". *Proceedings of the 11th international conference on World Wide Web WWW '02*, May 7–11, 2002, Honolulu, Hawaii, USA. ACM1-58113-449-5/02/0005.
- Rajashree Shettar, Dr. Shobha G, "Web Crawler On Client Machine", *Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol II IMECS 2008*, 19-21 March, 2008, Hong Kong.
- Eytan Adar, Jaime Teevan, Susan T. Dumais and Jonathan L. Elsas "The Web Changes Everything: Understanding the Dynamics of Web Content", *ACM 2009*.
- A. K. Sharma, J.P. Gupta and D. P. Agarwal "PARCHED: An Architecture of a Parallel Crawler based on Augmented Hypertext Documents", *International Journal of Advancements in Technology*, pp. 270-283, October 2010.
- Ashutosh Dixit and Dr. A. K. Sharma, "A Mathematical Model for Crawler Revisit Frequency", *IEEE 2nd International Advance Computing Conference*, pp. 316-319, 2010.
- Shruti Sharma, A.K.Sharma and J.P.Gupta "A Novel Architecture Of a Parallel Web Crawler", *International Journal of Computer Applications (0975 - 8887) Volume 14- No.4*, pp. 38-42, January 2011.
- Alex Goh, KwangLeng, Ravi Kumar P, Ashutosh Kumar Singh and Rajendra Kumar Dash "PyBot: An Algorithm for Web Crawling",

- IEEE 2011.
- Song Zheng, "Genetic and Ant Algorithms Based Focused Crawler Design", Second International Conference on Innovations in Bio-inspired Computing and Applications pp. 374-378, 2011.
- Lili Yana, ZhanjiGuia, Wencai Dub and QingjuGuoa "An Improved PageRank Method based on Genetic Algorithm for Web Search", *Procedia Engineering*, pp. 2983-2987, Elsevier 2011.
- AnduenaBalla, Athena Stassopoulou and Marios D. Dikaiakos (2011), "Real-time Web Crawler Detection", 18th International Conference on Telecommunications, pp. 428-432, 2011.
- BahadorSaket and FarnazBehrang "A New Crawling Method Based on AntNet Genetic and Routing Algorithms", International Symposium on Computing, Communication, and Control, pp. 350-355, IACSIT Press, Singapore, 2011.
- Anbukodi.S and MuthuManickam.K "Reducing Web Crawler Overhead using Mobile Crawler", *PROCEEDINGS OF ICETECT*, pp. 926-932, 2011.
- K. S. Kim, K. Y. Kim, K. H. Lee, T. K. Kim, and W. S. Cho "Design and Implementation of Web Crawler Based on Dynamic Web Collection Cycle", pp. 562-566, IEEE 2012.
- MetaCrawler Search Engine, available at: <http://www.metacrawler.com>.
- Cho, J. and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. *Vldb '00*, 200-209, 2000.
- Douglis, F., A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: A live study of the World Wide Web. *USENIX Symposium on Internet Technologies and Systems*, 1997.
- Fetterly, D., M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of Web pages. *WWW '03*, 669-678, 2003.
- Kim, J. K., and S. H. Lee. An empirical study of the change of Web pages. *APWeb '05*, 632-642, 2005.
- Koehler, W. Web page change and persistence: A four year longitudinal study. *JASIST*, 53(2), 162-171, 2002.
- Kwon, S. H., S. H. Lee, and S. J. Kim. Effective criteria for Web page changes. In *Proceedings of APWeb '06*, 837-842, 2006.
- Ntoulas, A., Cho, J., and Olston, C. What's new on the Web? The evolution of the Web from a search engine perspective. *WWW '04*, 1-12, 2004.
- Olston, C. and Pandey, S. Recrawl scheduling based on information longevity. *WWW '08*, 437-446, 2008.
- Pitkow, J. and Pirolli, P. Life, death, and lawfulness on the electronic frontier. *CHI '97*, 383-390, 1997.
- Selberg, E. and Etzioni, O. On the instability of Web search engines. In *Proceedings of RIAO '00*, 2000.
- Teevan, J., E. Adar, R. Jones, and M. A. Potts. Information retrieval: repeat queries in Yahoo's logs. *SIGIR '07*, 151-158, 2007.
- <https://www.javatpoint.com/thread-sleep-in-java>
- http://edukacja.3bird.pl/grafika/informatyka/python_institute/python-institute-certificate-pcep-30-01_sJsR.7kWb.FZ5h.png
- <https://www.pdf-archive.com/2020/08/27/pythoncertificate/prview-pythoncertificate-1.jpg>
- <https://averagelinuxuser.com/make-a-bootable-usb-drive-in-linux/>
- http://www.java2s.com/Tutorials/Java/OCA_Mock_Exam_Questions_2/Q5-1.htm
- <https://www.lifewire.com/what-is-web-directory-3482036>
- <https://www.computerhope.com/jargon/s/searengi.htm>
- <https://www.britannica.com/story/whats-the-difference-between-the-deep-web-and-the-dark-web>
- https://en.wikipedia.org/wiki/Distributed_web_crawling
- <https://gitmind.com/architecture-diagram.html>
- <https://www.edrawsoft.com/what-is-uml-diagram.html>
- <https://opensource.com/resources/linux>
- <https://www.pythonforbeginners.com/learn-python/what-is-python>
- <https://www.torproject.org/>
- <https://docs.python.org/3/library/re.html>
- <https://docs.python.org/3/library/os.html>
- <https://docs.python.org/3/library/sys.html>
- https://www.tutorialspoint.com/python_network_programming/python_http_requests.htm
- <https://pypi.org/project/PySocks/>
- <https://realpython.com/python-sockets/>
- <https://docs.python.org/3/library/argparse.html>
- <https://pythongeeks.org/subprocess-in-python/>
- <https://www.geeksforgeeks.org/python-urllib-module/>
- <https://www.pythonforbeginners.com/beautifulsoup/beautifulsoup-4-python>
- <https://www.geeksforgeeks.org/platform-module-in-python/>