RESEARCH ARTICLE OPEN ACCESS

A Methodical Study of Web Crawler

Vandana Shrivastava
Assistant Professor, S.S. Jain Subodh P.G. (Autonomous) College Jaipur, Research Scholar, Jaipur National University, Jaipur

ABSTRACT
World Wide Web (or simply web) is a massive, wealthy, preferable, effortlessly available and appropriate source of information and its users are increasing very swiftly now a day. To salvage information from web, search engines are used which access web pages as per the requirement of the users. The size of the web is very wide and contains structured, semi structured and unstructured data. Most of the data present in the web is unmanaged so it is not possible to access the whole web at once in a single attempt, so search engine use web crawler. Web crawler is a vital part of the search engine. It is a program that navigates the web and downloads the references of the web pages. Search engine runs several instances of the crawlers on wide spread servers to get diversified information from them. The web crawler crawls from one page to another in the World Wide Web, fetch the webpage, load the content of the page to search engine's database and index it. Index is a huge database of words and text that occur on different webpage. This paper presents a systematic study of the web crawler. The study of web crawler is very important because properly designed web crawlers always yield well results most of the time.
Key Words- database, search engine, Uniform Resource Locator, Web Crawler, web repository, website, world wide web

---

Date Of Submission:18-10-2018 Date Of Acceptance: 04-11-2018

I. INTRODUCTION

Web crawler also known as spider or web robot is a program that automatically traverses the large numbers of web pages by following hyperlinks, index them and stores the traversed web pages links for prospect use. Web crawler is a program that downloads the web pages automatically from huge and gigantic web. Major part of the data present in the web is unstructured data that is not organized in a specified way and has no predefined data model. This data generally contains more text data which may have variety of dissimilar formats. Major task of the crawler is to maintain and manages the index the web pages and make searching fast, truthful and productive. Pages visited by the web crawlers are copied in web repository for later use. Crawler is a tool to collect and keep the database up-to-date.

As the contents and users of World Wide Web is increasing very swiftly day by day so it is very essential to search the information, links and data properly in shortest span of time to dig out most pertinent information within expected time. Web crawlers are used by search engines to serve the same purpose. A lot of search is being done to devise the perfect spider (web crawler) so that the searching and crawling process can be improved to a huge extent. Different search engines use different types of web crawlers for example Google uses GoogleBot, Microsoft's web crawler is Msnbot/Bingbot, Baidu search engine uses Baidu Spider. All web crawlers are supposed to follow some rules and regulations and work on agreed standards, Google and Bing are working on standards.

The key working of all web robots is same yet there is some difference in the crawling process and algorithms they used for searching and retrieving information from web. Whenever searching is carried out by users, they may not get the links and pages that are completely useful. The most challenging task for users is to get an appropriate search engine that can provide relevant, robust and optimal number of pages during searching. Currently the most popular search engine is Google used by 74.54% of internet users all around the world and other like AOL, Ask and Duckduckgo has less that 1% market share.

II. REVIEW ON WEB CRAWLER

Different research papers have been studied on search engines, web crawler and crawling techniques. These research papers are helpful in analyzing the present work done and detecting the lacunas which remains unsolved in the current work.

Mridul B. Sahu et al. in their paper "A Survey on Various Kinds of Web Crawlers and

Intelligent Crawler" explained in the paper that two major approaches control the decisions made by the crawler [1]. First approach known as supervised learning decides its crawling strategy by looking for the next best link amongst all links it can travel whereas the second approach computes the benefit of traveling to all links and ranks them, which is used to decide the next link. Efficiency upgrading can also be implemented by using data mining algorithms. They suggested that through learning procedure, a crawler is able to make intelligent decisions while selecting its strategy.

Dhiraj Khurana, Satish Kumar in paper "Web Crawler: A Review" explained various types of web crawlers and explained the pros and cons of each types of crawler. They also explained distributed crawler. In this crawling, multiple crawlers are being managed by the URL server which download web pages in parallel mode, the crawlers then send the downloaded pages to a central indexer on which links are extracted and sent via the URL server to the crawlers. This distributed nature is of great use as it not only reduces the hardware requirements and also increases the overall download rate and consistency.

Hetal J. Thankil et al. in paper "Domain Specific Web Crawler: A Survey" described that different crawlers use different crawling techniques but to improve quality of service (QoS) hybrid crawlers can be designed which will overcome the limitation of algorithm to improve the quality of crawler.[2]

Dr Jatinder Singh Bal et al. in the paper "Web Crawler: Extracting the Web Data" concluded that building an effective web crawler to solve different purposes is not a difficult task, but to choose right plan and built an effectual structure and its execution is actual challenge.[3] A number of crawling algorithms are used by the search engines. A good crawling algorithm should be implemented for improved results and high performance.

Trupti V. Udapure et al. in their paper "Study of Web Crawler and its Different Types" showed that web Crawler is the essential source of information retrieval which traverses the Web and downloads web documents that suit the user's need. Web crawler is used by the search engine and other users to regularly ensure that their database is up-to-date. The overview of different crawling technologies has been presented in this paper.[4]

Beena Mahar and C K Jha in their paper "A Comparative Study on Web Crawling for searching Hidden Web" described different crawling technologies and explained how to crawl hidden web documents with different ways.[5]

III. CHARACTERISTICS OF WEB CRAWLER

As the Web Crawler is essential part of the Search Engine it must have following characteristics-
Efficient- Processor, bandwidth, memory and storage are valuable resources of the system so a web crawler should be elegant to use them effectively and efficiently.
Scalable- A web crawler must have scalable so that it can be add on more machines and extends the bandwidth.
Robust- Web consists not only static pages but is also contains dynamic pages as well, so crawler must be compatible with static as well as dynamic pages.
Extensible-Web crawler should be extensible enough to manage with new data structures and new protocols.
Quality Content- Quality and meaningful contents must be identified by the crawler and index them on priority basis.
Duplicate Content- Crawler must be able to distinguish and remove duplicate data obtainable on various websites. Various methods like visitor counter, checksum can be used for content checking.
Distributed- Web crawler must be capable to execute on several machines concurrently.
Excluded Content- Most of the site has robot.txt file which has necessary information about how much file can be crawled and which part in prohibited from crawling so proper decision can be made for either visiting the site or not.
Spam Elimination-Blacklist URLs can lower down the priorities of the linked pages hence it is very necessary that crawler should identify and reject such links during crawling. [7]

IV. POLICIES USED BY WEB CRAWLERS

The operation and performance of a web crawler is the result of blend of various policies used by them to make search more significant, meaningful, precise and appropriate. These policies or guidelines are discussed here-
Assortment or Selection Policy- Internet holds huge data that contains hyperlinks, text, graphics, images, videos and many more formats & types. The objective of the users during searching is to download most relevant data among achieved links. The selection policy ensures that most pertinent pages would be downloaded among available large number of pages. The selection of pages and links are managed here so that searching is done for only useful outcome.

Revisit Policy-Every day new websites are uploaded, existing websites, webpage and their

content updated very regularly. Regular updation is an essential activity of web developers and designers to make the websites more informative with latest data. Some new pages are added, unnecessary pages are deleted and some matter of the webpage is changed. To make retrieval more significant the crawlers used revisit policy. This policy helps the crawlers to decide when they verify that pages are updated and links should get updated. Precisely this policy helps the spiders to plan their next visit to webpage to get latest data.

Politeness Policy- Web crawlers or robots can gain data very rapidly as they have crippling impact on the recital of a website. Web contains millions of pages and when a query is passed to search engines, lots of links can be positioned by the spiders. It is not possible to download all links at a glance so politeness policy is used to keep search more meaningful, quick and precise so to avoid overloading. This policy relinquish burden from the crawler to access all links in single effort.

Parallelization Policy- Dissimilar search engines use different web crawlers which run in parallel form on various sites to access links and webpage. An ideal Coordination is needed among them so that repeated download can be avoided from the same site. It also ensures crawling more and more sites in lesser stabs.

V. GENERAL ARCHITECTURE OF WEB CRAWLER

Web Crawler is an essential component of the search engine. It is the core part which is responsible to search websites and create index of traversed webpage. The speed of the web crawlers is very high to access several websites in fraction of time. It starts its searching with few seed URLs, traverses the websites and updates the search engine database very regularly and rapidly.

Web crawlers are used by search engines for the following reasons-

Crawler sustains mirror sites for popular web sites for speedy access of webpage and links.

It creates an index of crawled or accessed links.

Before downloading pages and links crawlers check their syntax and structure that it is valid or not.

It keeps a close watch on websites that when the content or composition of the same get changed so that the search engine database can be updated accordingly.

It can also be used to look for copyright or exclusive rights violation.

Following figure shows the architecture of a typical Web crawler.

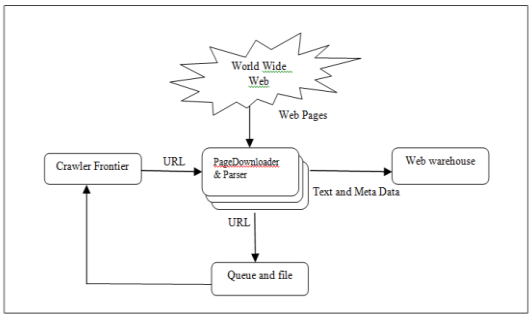

Fig 1.1 Web Crawler

When a page is traversed by the crawler it starts with the available URL (Uniform Resource Locator) known as seed or kernel .When URL is visited, different hyperlink accessible to that page is added in list of URLs. This list is known as Crawler Edge or Frontier. Different URLs and hyperlinks from visited websites are downloaded and stored in the form of index in database of search engine. Accessed hyperlinks are used to extract webpage from web and stored by the page downloader and parser in the data warehouse, which contains index, hyperlink, webpage, metadata, keywords and phrase.

Web Crawler uses various algorithms to access suitable link from index to find suitable webpage for users. Crawler itself does not access each page of the website but it sends request to web servers for providing links and webpage. Links of lots of pages are visited by the crawler in a fraction of time and the database is updated regularly.

Crawling Process –Web crawler is the central part of any search engine. Its main task is to search, navigate and download the references and content of the visited web pages in search engine database. When crawler creeps the website, it chases every hyperlink of each visited web page and collects the documents and a searchable index is built for each visited page on the search engine's. Search engine database repository is the place where contents are loaded. Most of the website has its own robots.txt file.[6] It's an important file which holds information about which part of the website to catalog and which part be overlook. This file contains information about prohibiting instruction as some users don't want that a specific part of the website is being traversed, so rotots.txt is a controlling tool for spiders. A single search engine may scuttle thousands of instances of web crawling on manifold servers to build its huge database.[6]

A general methodology adopted for the web crawling is depicted by following steps-

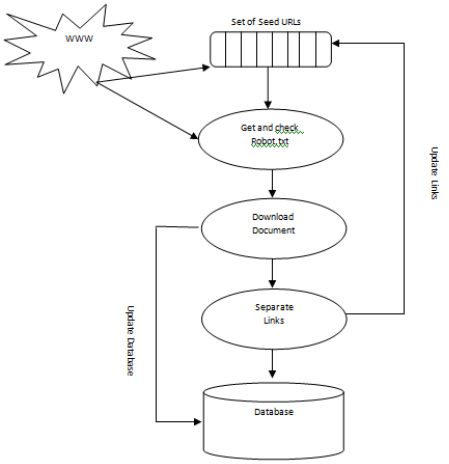

Fig 1.2 Working of Web Crawler

Whenever a text in the form of query is passed to the search engine, web crawler crawls from set of seed URLs, group of crawlers engaged to observe the query and get subset of URLs to crawl.

Page downloader is used to fetch the webpage.

With new URLs, downloaded page is passed to extractor which starts extracting the words and contents of the webpage.

Crawler Frontier starts indexing by retrieving all the links and put the visited URLs in the web database.

It inspect robot.txt file (if available) if some prohibiting instruction are there or not.

Take out different words, keywords and phrases in the database for future search.

Abstract and visited date of the traversed page is saved so that system has complete detail about when to check again that page later on.

For each link visited repeats the process for retrieving the new pages added (if any).

When the spider doesn't find a page, it will eventually be deleted from the index. However, some of the spiders will check again for a second time to verify that the page really is offline.

Filtering content of webpage is optional part for crawler, some perform it and some just ignore it.

VI. TYPES OF WEB CRAWLER

A. Customary Web Crawler-This is also known as Traditional Web Crawler. Each website has its URL contains multiple web pages. These web crawlers used multipage interface method for accessing web pages. Here individual page of the website is treated separately and having a unique URL known as seed or kernel URL.

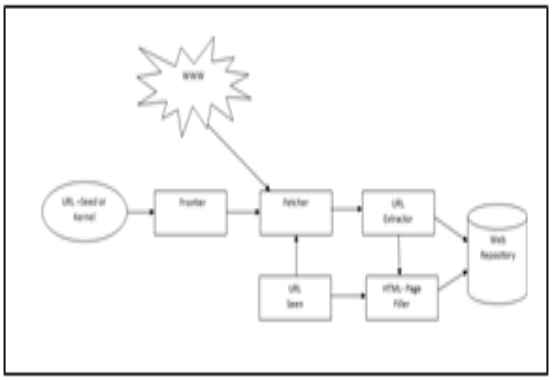

Fig 1.3 Architecture of Traditional Web Crawler

When an input seed URL from as deposit of URL is provided to the frontier, it is passed to the fetcher. Fetcher is responsible to access the web contents related to specific URL. Web data, hyperlinks and many more data is traversed by the fetcher and passed to the URL extractor. The major task of the Extractor is to take out new links from parsed web pages and the new links are handover to the page filler module and web repository. Web repository interacts with data files and checked for the new links. If some new links are not of use of web crawler, the page filler filters out the unwanted URLs and keeps only useful URLs. These URLs are passed to the net URL-seen module which check whether this URL is visited, if not it is passed to the fetcher and the cycle continues until all accessible links are visited.

B. Deep Web Crawler-There is millions of websites available on the net and plenty of webpages are accessed daily by the crawlers. This is only 2-4% of visible data available to the users which is also known as publically indexable web (PIW) but more than 90% of data are not even accessible by the crawlers, this unaccessed data is known hidden data, deep web or hidden web consist of millions of web pages. Generally web robot accesses the data or set of pages which can be accessed by simply following the hyperlinks. If the pages or links demand for authorization or previous registration, then such pages are ignore by them but it consists of lots of quality web data hidden behind search forms. To retrieve webpage and links from hidden data special crawlers are designed called deep web crawlers.

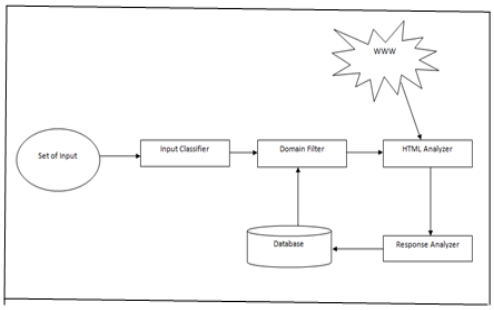

Fig 1.4 Architecture of Deep Web Crawler

Deep or profound web crawler consists of following major module-Input Classifier- It is responsible for checking the relevance of downloaded web page with the input string. When an input set of seed URLs, domain data and client precise are supplied to it, it selects the HTML components to communicate with input set.
Domain Filter-It uses the input data to fill the HTML form and the result is being passed to the HTML analyzer for further action.
HTML Analyzer-It submits the filled form to the web server and new web pages are obtained from web server.
Response Analyzer-New web pages are checked by the response analyzer and added to the database and this whole process run continues.
Following sequence is followed by HiWE (hidden web explorer)-
Form Analysis-Check the form syntactically and process it filling.
Value Assignment-Data submitted by filling the form is matched with the data of web server by matcher to generate a set of value assignment for ranking purpose. Usually a fuzzy aggregation function is used for value assignment
Form Submission-The filled form is submitted to respective URLs.
Response Analysis-The page received from the web server is analyzed.

C. Incremental web crawler-An important need of web repository is to keep its contents fresh and purge the old or outdated pages to keep space for new pages. Incremental crawler updates available set of pages and keep their fresh copy. To accomplish the task, this crawler does not start the crawling from scratch but keep track whether already available pages have changed since last time it was creep. It frequently visits the web to keep the web repository contents fresh and newest and next visit to web is decided as how often the page has changed. [6]

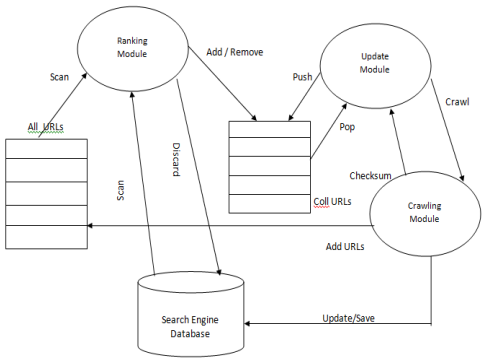

Fig 1.5 Architecture of Incremental web Crawler

Incremental Crawler has following modules-
a) Ranking Module-This module is responsible to rank the crawled pages and performs the modification decision. When web is accessed, this module decides that the crawled page should be added to the URLs or not. When new crawled webpage which is not present in the database is being added, then it is the duty of ranking module to remove unnecessary page and make home for new page by applying proper replacement policy. The URL list is also updated with immediate effect so that update module can get it immediately.
b) Update-Module- When a URL is added to the web repository, it is the task of update module to revise the existing web page by their fresh copy (if it is changed).Importance and estimated change frequency of page are two major factors which decide when to revisit the page. Frequently changed pages are kept on the head of URLs queue so that they can be visited frequently. This module maintains the freshness.
c) Crawl-Module- This module simply accesses the web and updates the URLs list. It also helps the update module whether a page is being pushed or popped from web data.
Incremental crawler serves the following purposes-
a) Revisit the web to crawl pages for Data enrichment
b) Keeps the existing content fresh and improve quality of local dataset
c) Save network bandwidth
d) Make space for new pages

D. RIA (Rich Internet Application) Web Crawler
Web is a huge, preferable and suitable source to access information and its applications are more interactive, suitable, approachable and responsive for users. These applications are Rich Internet Application. Ajax and Dom technologies are used in RIA application to access data but both are very time consuming. To condense the access time, special crawlers known as RIA crawlers are used. Crawljax is the example of RIA crawler that seizes

the Ajax state and then decides for testing and indexing. This is a unique web crawler which works in different manner than traditional crawler, it primarily judges the user interface and changes made by users are taken into account to transmit the crawling strategy. During searching, the JS Engine starts with the web browser, runs it and access the initial web page linked with URL seed and puts it in the browser. Constructed DOM is given to DOM state to know that it is seen for first time or not. First time occurred DOM sate is passed to the extractor to extort the JS event, selected JS event is passed to strategy module and then passed to the JS engine for more implementation. The process continues until next accessible DOM states are seen.

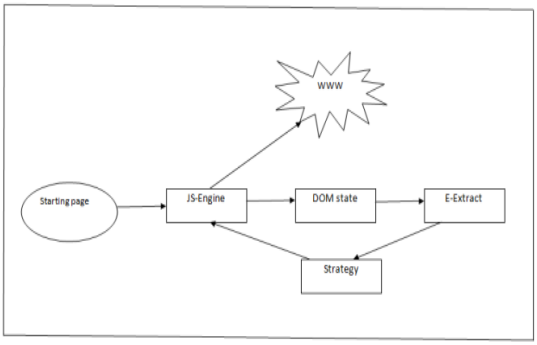

Fig 1.6 Architecture of RIA Crawler

E. Unified model - Web Crawler

UML or Unified Modeling Language used by software developers for software system, business modeling and non software systems. In UML web browser, Document Object Model (DOM) and Uniform Resource Locator (URL) are two basic factors on which the node (page) is calculated. Browser redirection is the is a special client side event and user model consists of following three major following module –

a) User Module- It consists of use case diagram
b) Object Module-It represent the class diagram
c) Dynamic Module-It signify the sequence diagram

F. Focused Web Crawler-Also known as Topical or Topic Oriented Crawler. Search engine database contains plenty of data and information within it. Selecting the most relevant links and pages related to specified topic is probably the most challenging task. In today's scenario, users are more specific towards topic and its relevance. Special crawlers which are topic or theme specific are known as focused crawler. The biggest advantage of focused crawler is that is reduces the network traffic and download only relevant pages. Prior to downloading, this crawler foresees that a link is relevant to the topic or not and then only downloads the pages based on query.

When a searching started with a keyword or phrase, generally crawler starts it from the highly ranked pages and immediate topic oriented but less ranked pages remains unaccessed many times. These crawler recursively perform a depth access from high ranked pages to lower ranked pages to get topic oriented contents rather than perform only surface search.

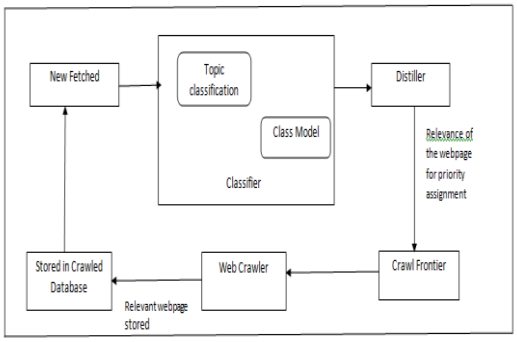

Fig1.7 Focused Web Crawler

Focused web crawler has three important parts-

a) Classifier-It is responsible for making significance conclusion on crawled web pages to decide which link should be expanded and get indexed in index. It also partitioned the web contents in hierarchical order according to different categories. It can follow hard focusing or soft focusing rule.

b) Distiller- Apart from relevance another main task is to filter the pages that contains elite set of hyperlinks follow more and more pages as per topic. The distiller follows hyperlinks to rich contents of web repository and measures the visit precedence of downloaded web pages on priority basis.

c) Crawler- Its search task is dynamically reconfigurable under the supervision of distiller and classifier and priorities are also determined by them only. The crawler searches pages, make index, explore the links and store the data in web repository with multiple threads architecture. Memory is shared among multiple threads and new crawling tasks are assigned to them.

G. **Parallel Crawler**-Web is gigantic contains millions of webpage, hyperlinks and other information and it is almost impossible to access and retrieve all or noteworthy part of the web in a single effort. Therefore many search engines often run several procedures in equivalent manner to exploit more and more page, links and significant part of the web to rich search engine database, such type of crawlers are known as Parallel crawler. Here multiple crawling processes are accomplished

in parallel manner. The downloaded web pages are stored locally and URLs are extorted and links are used to follow the webpage. These crawlers are very useful to scatter the network load over several networks as heavy load on a single network is not feasible and can't be handled so overall network load is dispersed by dividing the areas to be creep by each instance of the web crawler.

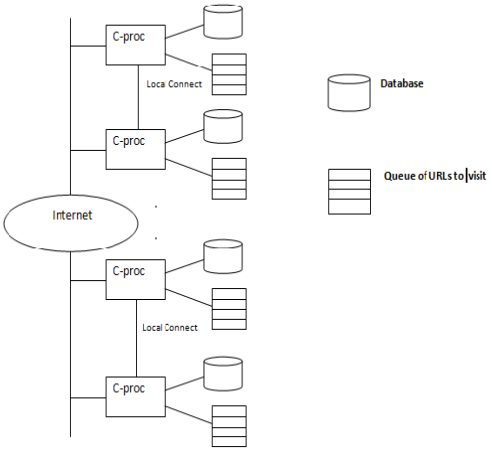

Fig1.8 Architecture of Parallel Crawler

These crawlers may be of two types based on the location of traversing-
a) Inter site- When several crawling processes are geographically dispersed, use different network when download pages from far-off sites and run on different networks it is inter site parallel crawler.
b) Intra site- When several crawling processes run on local network and connected over same LAN to download the web pages it is intra site parallel crawler. [6]

H. Distributed Web Crawler-Indexing is the most challenging task for crawler, as the size of the web grows at fiery rate, it is very essential to complete the crawling task in rational time. To extract more and more data in lesser time, it is needed to parallelize the crawling process and must be distributed to several processes to make it more scalable. This goal can be achieved by Distributed Crawlers that uses dispersed computing technique.

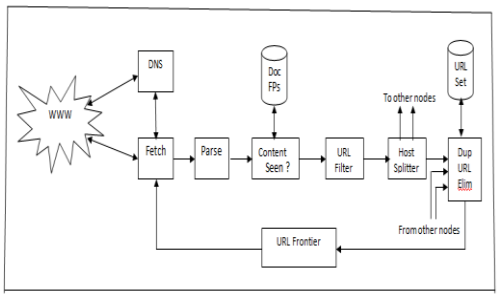

Fig 1.9 Architecture of Distributed crawler

Here URL server dispenses separate URLs to many crawlers so that downloading can be accomplished in parallel mode. These downloaded pages now send to core indexer where hyperlinks are digged up and send back to URL server to the crawlers. The nature of distributed crawler not only reduces hardware requirement but also increases the download speed and reliability[5].

VII. CONCLUSION

World Wide Web is a huge repository of millions of web pages and data related to dissimilar topics. It has no centrally organized content arrangement. Hence searching data related to a specific topic is not an easy task. With such challenging tasks the role of web crawler become more important and its design issues must be taken into consideration to produce more effective results. This paper presents a detailed study of web crawler, its architecture and types.

VIII. FUTURE WORK

Web database is a preferred and easily available resource of information for professionals, researchers, innovators and users. To provide full advantage of web database, it is necessary to work with more new ideas and new crawler design. As more information will be available more progressive growth can be implemented in new design and techniques of crawlers. Hence to maintain the Quality of Service, fast and relevant accessing, reduced response time and improve productivity it is necessary to design a crawler that can be a hybrid one and serve diversified purpose by applying threading and rapid text mining using AI tools.

REFERENCES

[1]. Mridul B.Sahu, Prof. Bharne., Samiksha (2016) A Survey on Various Kinds Of Web Crawlers And Intelligent Crawler International Journal of Scientific Engineering and Applied Science ISSN: 2395-3470 Vol-2, Issue-3,March 2016
[2]. Dheeraj Khurana,SatishKumar(2012) Web Crawler: A Review International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012 ISSN (Online): 2231 –5268
[3]. Thankil J. Hetal, PoojaSawant, NishaKasale, SonamYerawar , Prasad Belhe (2015) Domain Specific Web Crawler: A Survey International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 12, December 2015
[4]. Trupti V. Udapure , Ravindra D. Kale , Rajesh C. Dharmik (2014) Study of Web

Crawler and its Different Types IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05

[5]. BeenaMahar ,Jha, C K (2015) A comparative Study on Web Crawling for searching Hidden Web International Journal of Computer Science and Information Technologies, ISSN : 0975-9646 Vol. 6 (3) ,pp. 2159-2163

[6]. Shruti Sharma, Parul Gupta (2015) The Anatomy of Web Crawlers International Conference on Computing, Communication and Automation ISBN:978-1-4799-8890-7/15/ IEEE 849

[7]. M. Ahuja,J.P.S.Bal, Varnica (2014) "Web Crawler: Extracting the Web Data"I nternational Journal of Computer Trends and Technology – ISSN: 2231-2803,13,pp.132

[8]. Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, and Hai Jin(2016) Smart Crawler: A Two-Stage Crawler for Efficiently Harvesting Deep-Web Interfaces IEEE Transactions On Services Computing, Vol. 9, No. 4, July/August 2016

[9]. Aviral Nigam (2014) Web Crawling Algorithm International Journal of Computer Science and Artificial Intelligence International Journal of Computer Science and Artificial Intelligence Vol. 4 Iss. 3, PP. 63-67

[10]. Subhendu kumarpani, Deepak Mohapatra, Bikram Keshari Ratha (2010), Integration of Web mining and web crawler: Relevance and State of Art, International Journal on Computer Science and Engineering Vol. 02, No. 03, 772-776

[11]. Nemeslaki, András; Pocsarovszky, Károly (2011), Web crawler research methodology,, 22nd European Regional Conference of the International Telecommunications Society

Vandana Shrivastava "A Methodical Study Of Web Crawler "'International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.11, 2018, pp 01-08