International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

Different kinds of Web Crawler

Vishakha
Department of Computer Engineering
Shri Ram College of Engineering and Management
Faridabad, Haryana, India
vishumrce@gmail.com

Abstract **- A Web crawler is an Internet bot which helps in indexing the web. It crawls one page at a time through a website until all pages has been indexed. Web crawler helps in collecting information about a website and the links related to them. A web crawler is defined as mechanized program that accurately scans through Internet pages and downloads any page that can be reached by links A Web crawler is also known as a Web spider, automatic indexer or simply crawler. Nowadays finding the relevant information over millions bytes of data is very difficult. For this task we need to use efficient crawler so that it can give us best result within minimum time. This paper basically focused on the discussing different crawlers with their efficiency.**

Keywords- **Web Spider, Automatic Indexer, Scans.**

I. INTRODUCTION

The internet has becoming the largest unstructured database for accessing information over the documents day by day. [1] It is well known that the information technology has a immense effect on the conduct of the business activities, and the Internet is now becoming the largest marketplace of the world. Innovative business professionals have developed the commercial applications of the Internet according to the need of their customers. [2] With the increasing amount of electronic text from the complex the World Wide Web, more and more knowledge you need is included on the daily basis. But, the huge amount of text also takes so much trouble to people to find useful information.

For example, the standard Web search engines have very low precision, since typically some Web pages returned resulted information mixed with a large number of irrelevant pages, which is due to the situation that the topic-specific features may occur in different contexts. So, one suitable way of organizing this vast amount of documents is necessary.[3] The World Wide Web is an architectural structure for accessing linked documents spread over millions of websites all over the Internet. The visible Web with its estimated size of at least 20 billion pages causes a challenge of retrieval of relevant information.

Even with increasing hardware and bandwidth resources at their disposal, it is very difficult or almost impossible for search engines to keep up with the growth of the Web [4 The retrieval challenges further compounded by the fact that Web pages also change very frequently. Thus, inspite of all the attempts of search engines to index the whole Web efficiently, it is expected that the subspace eluding indexing will continue to keep grow. Therefore, retrieval of relevant information from the Web has been considered one of the most important strategies to build digital libraries capable of covering the whole we band take benefit from the large amount of information and resources that can be useful.

1. Background

The World Wide Web gives us a huge source of information of almost all type we want. However this information is not provided at a single site else distributed among many web servers and hosts, representing in different kinds of formats. We all want that we should get the best possible search result in less time. For every crawler there are two issues that it have to consider always. Firstly, The web crawler should have the capability to plan where the plan is to decide which pages are going to be downloaded next. Secondly, It needs to have a well optimized and robust system architecture so that it can download a huge number of pages per second even against crashes, manageable, and also take care of resources and web servers.

II.FEATURES OF A CRAWLER

1. Robustness This is the capability of a web crawler to handle spider traps which generate web pages that deceive crawlers so that they get stuck by retrieving a countless number of pages in a specific domain. So the crawlers must be designed to be durable enough to such spider traps. It may also happen that not all of the spider traps are harmful; some are the side-effects of an erroneous website development.

2. Politeness The speed at which the crawlers visit the website. The crawler must only retrieve one page at a single time from the same site and should wait for some time between 2 successive download. Also crawlers must only crawls the websites that are allowed to crawl. The consent to crawl a website is given by robots.txt file.

© 2019 IJSRET

3. Distribution- The crawling should be distributed within several machines. This is to acquire advantages in deployment, programming, and debugging.
4. Scalability- The architecture of the crawler should allow balancing of the crawl rate by inserting up further machines and bandwidth.
5. Efficiency- This is the smart use of the memory, bandwidth and processor. E.g- as few idle processes as possible.
6. Quality- The web crawler should detect the most useful pages, to be indexed first. Simply the pages of high quality or highest page rank should be crawled first.
7. Freshness- The web crawler should constantly crawl the web so that it can update the information. The database should consist of latest pages.
8. Extensibility- A web crawler should be extensible to handle a new data structure like XML/EXML, new protocols etc.

III. ARCHITECTURE OF WEB CRAWLER

The web search engines have a very important component this is known as web crawler. Today the growth of web crawler has to be done according to the updated web. A list of URLs is available with the every web crawler and each URL is called a seed. Each URL has to be visited by the web crawler very efficiently . It identifies the different hyperlinks present in the page and updated the list by adding those hyperlinks to the list of URLs to visit. This list is called crawl frontier.

The URLs in the crawl frontier are always visited individually By Using a set of rules and policies. Different pages from the internet are download by the parser and the generator and stored in the database system of the search engine. After this, queue has to be made and placed all the URLs in it and later gets scheduled by the scheduler and can be accessed one by one by the search engine whenever required. The links and related files that has to be searched can be made available whenever needed next time according to the necessities. With the help of suitable algorithms web crawlers find the all related links on the pages for the search engines and use them further. Databases are very big machines like DB2, used to store large amount of data [5].

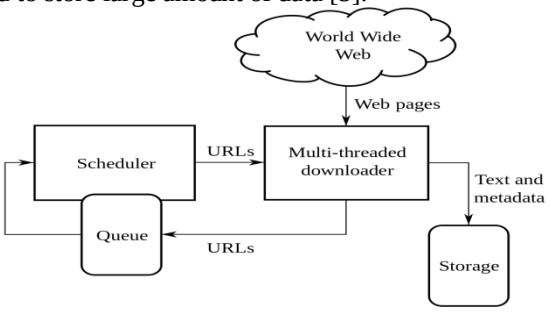

Fig. 1 Architecture of Web Crawler.

IV.TYPES OF WEB CRAWLERS

1. Breadth first crawler- Begins with a bunch of pages and then searches the remaining pages by following links in the breadth first manner[6].The frontier is employed as a First in First out (FIFO) queue. The process of crawling in breadth first fashion is:

- All the given seeds are kept in the queue.
- Lists of visited nodes are made which are at first empty.
- If the queue isn't vacant:- o the initial node from the queue is discarded othat node is appended to the list of visited nodes
- for each of the edge beginning at the node:-
- Does nothing with that edge if the node at the last is in the queue already or occurs on the list of visited nodes.
- Else, the node is appended at the last of the edge up to the end of the queue. In reality web pages are not navigated strictly in breadth first manner but it may use a variety of policies. For example it may crawl most significant pages first.

2. Incremental web crawler- Incremental crawler [7] updates current group of downloaded pages as an alternative of restarting the crawl from the beginning each time. This engrosses a way of ascertainment if a page has changed from the last time it was crawled. A crawler will persistently crawl the whole web supported by many group of crawling cycles. An adaptive model is used which uses data from preceding cycles to determine which pages should be tested for updates resulting in huge freshness and achievement of low peak load[8]. So, if the crawler can estimated how frequently pages changes, then the incremental web crawler may re-visit only those pages that have been changed rather than reviving the whole collection together.

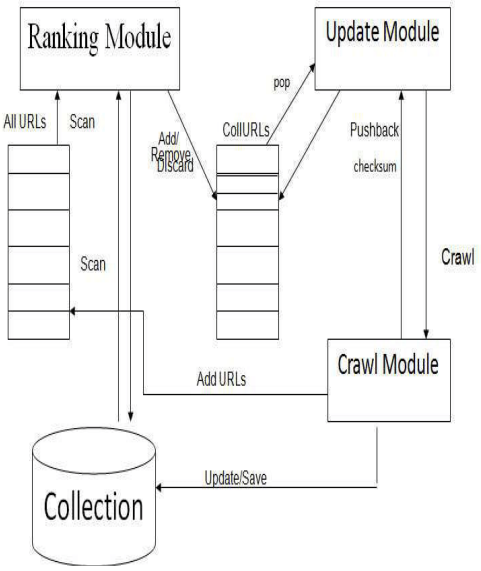

Fig 2. Architecture of Incremental Web Crawler.

© 2019 IJSRET

3. Focused web crawler- Focused crawler [9] is such web crawler that downloads the web pages which are correlated with each other. It collects documents which are definite and correlated to the given topic. Therefore, a focused crawler looks for, obtains, indexes and keep up pages on definite set of topics which represents a relevantly narrow segment of the web. It is also called as theTopic Crawler for the cause that is its way of working. The focused crawler ascertains the Relevancy and Way Forward.
It ascertains to how much extent the given page is related to the particular topic and how to proceed forward [10] The advantages of focused web crawler are-

- It is cost-effectively in terms of both hardware and network resources.
- It can also reduce the network traffic amount and downloads on a website.

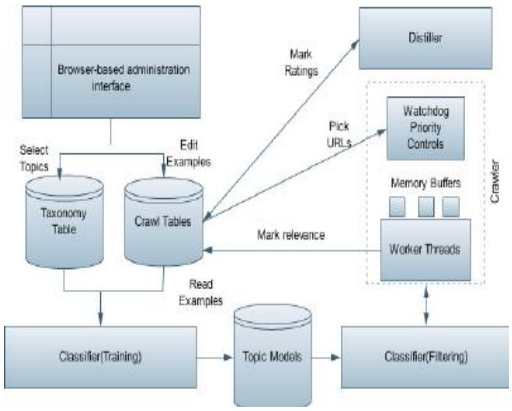

Fig. 3 Architecture of Focused Web Crawler.

4. Hidden web crawler- Here the information of web content is hide behind web forms and the client side scripting is referred to as the hidden web, which is estimated to consist of many millions of web pages. In the world wild web the huge amount of information is available on different websites. But there are some web pages which are not directly located by the search engines because today in almost all search engines searchable databases are not properly index able or query able. So they appear hidden to the internet user. such pages are referred to as the Hidden Web or the Deep Web.

The deep web is the largest growing area of information on the internet now days. In the hidden web crawler the Data on the web is located in the file/database and can be fetched by providing queries or by filling up the forms on the web. Current day crawlers crawl only publicly index-able web (PIW) That means set of pages which are accessible by following hyperlinks, overlooking search pages and forms which require authorization or earlier registration. Recent study is estimated that the hidden web's size is 500 times more than the size that is of the publicly index-able web (PIW).These web crawlers are usually implemented by central bureau, trademarks offices, etc.

3. Parallel web crawler- Because the growing amount of data on web, it becomes more difficult to retrieve the whole or a significant portion of the web from the million tons of data by using a single process. Therefore, lots of search engines frequently execute multiple processes in parallel to do the above task, so that download speed is taken full advantage. Multiple crawlers are frequently executed in parallel which are therefore known as Parallel crawlers . Consists of crawling processes called as C-procs. Problems associated with parallel crawling are:

1. Overlapping- It is sometimes possible that more than one c-proc might download the same content of information again and again when a multiple number of c-procs are running at a single time is known as overlapping.

2. Quality- More important pages are downloaded first, however this could be a difficult task for parallel web crawlers. Bandwidth of communication: Communication can develop importantly if the numbers of c-procs augment and must be reduced to preserve the effectiveness of the crawl.

6. Distributed web crawler- This crawler [11] sprints on group of workstations. In distributed web crawler, a Uniform Resource Locator (URL) server assigns peculiar URLs to multiple crawlers, which downloads web pages in parallel, the crawlers will then send the downloaded pages to an innermost indexer on which links are retrieved and sent by means of the URL server up to the crawlers. This distributed character of crawling process decreases the hardware requirements and increases the largely download speed and dependability. First distributed crawler was developed which uses client server model that is centralized. But with this model there were problems like obstruction, expensive administration and being a single point of failure. But when it is developed to be fully distributed crawler where there is no central controller because of which problems like scalability, increased autonomy of nodes and failure resilience have been overcome. Although, fault tolerance is considered, then if an error occurs, the whole system will be difficult to maintain, and may be very difficult to ignore the loss of information.

V. ADVANTAGES AND DISADVANTAGES OF WEB CRAWLER

1. Advantages- The web crawlers offer huge databases of websites for searching. The whole text of individual web pages is often searchable. Good for searching very vague terms or phrases.

2. Disadvantages- There is no need for humans to dig out the difficulties, such as duplicates. The large of database may lead to huge numbers of search results. The

search commands could be frequently be compound and baffling.

VI. APPLICATION AND RESEARCH AREA GENERAL WEB SEARCH

A web crawler for general search of web engine should carefully balance between coverage and quality. Coverage means that it must scan pages that allow it to create index which may be used to answer many difficult queries. Quality means that the pages must be of a high quality. Vertical crawling: Vertical crawling for data collection: It can be used to collect data from different sources. Most common type of this crawler is shopbot, that is intended to download information from on- line shopping lists and provides an interface for evaluating prices. News crawler to gather news items from sources that are pre-defined. Vertical crawling of specific formats: This search also includes segmentation by a data format. There are several machines that crawl image, audio, or videos.

1. Web characterization- It is the requirement for building effective web search engines. This is a difficult issue the even if the web contains a finite quantity of information, it may contain infinite pages. Page centered characterization measures page sizes, technologies, and other properties. While for link centered characterization the choice of starting URLs for performing the crawl is difficult.

2. Mirroring- It is the act of keeping a partial or complete copy of a web site. The aim of mirroring is to distribute server load and provide quicker access to users in distinct network locations. The copies are called mirrors. Web archiving does the act of keeping a huge set of pages without removing the outdated copies, i.e., the entire history of the each page is kept**.**

3. Web site analysis- A web crawler can be used to analyze a website and it can make changes on its own according to criteria that are predefined. The majority of the common form of this is link validation which automatically scans the web pages from broken links to inexistent pages. Another application is the web directory that automatically finds the web sites that are not available and to test them the attention of the directory's editors is called.

4. Semantic Based Crawling- The Automated ontologism formed in the paper[12] are based on lexical relations of a word i.e synonyms, concepts of hyponymy and hyponymy, meaning of the words. To get various, accurate details of important terms and the related terms these created ontology are used and semantic based web crawling can be easily done.

REFERENCES

[1]. S.Balan1, Dr.P.Ponmuthuramalingam2," A Study on Semantic Web Mining And Web Crawler", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 9 Sept., 2013 Page No. 2659-2662.

[2]. Abhiraj Darshakar,"Crawlers intelligence with Machine Learning and Data Mining integration", International Conference on Pervasive Computing (ICPC),2015.

[3]. Lu LIU, Tao PENG "Clustering-based topical Web crawling using CFu-tree guided by linkcontext" in Higher Education Press and SpringerVerlag Berlin Heidelberg 2014.

[4]. Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang "Ontology-Learning-Based Focused Crawling for Online Service Advertising Information Discovery and Classification" in Springer-Verlag Berlin Heidelberg 2012.

[5]. Mohit Malhotra (2013), "Web Crawler And It's Concepts".

[6]. Bing Liu, "Web Content Mining" the 14th international world wide web conference (www 2005) China .japan

[7]. Jumgoo Cho and Hector Garcia-Molina, "The evolution of the Web and implications for an incremental crawler", Prc. Of VLDB Conf, 2000.

[8]. Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli, Web Crawler in Mobile Systems, International Journal of Machine.

[9]. Paolo Boldi_ Bruno Codenotti† Massimo Santini‡ Sebastiano Vigna," UbiCrawler: A Scalable Fully Distributed Web Crawler".

[10]. Mahmud, Hasan, Moumie Soulemane, and Mohammad Rafiuzzaman. "A framework for dynamic indexing from hidden web." IJCSI (2011).

[11]. Vladislav Shkapenyuk and Torsten Suel, "Design and Implementation of a High performance Distributed Web Crawler", Technical Report, Department of Computer and Information Science, Polytechnic University, Brooklyn, July2001.

[12]. Preeti Kathiria ,Sukraat Ahluwalia, "A Naïve Method for Ontology Constructions" 2016, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016.

© 2019 IJSRET