ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

A Study on Web Crawler

Nikhil Chitre

B.E, Department of Information Technology, MAEER's MIT College of Engineering, Kothrud, Pune, Savitribai Phule Pune University, Pune, India

ABSTRACT: A web crawler which is also known as a spider or a robot is a computing program which crawls and browses the World Wide Web in an orderly and automated fashion. This process is known as Web spidering or crawling. Crawlers are mainly programmed to visit websites that have been put forward by their owners as new or updated. The entire sites or specific web pages can be selectively visited and then indexed. Crawlers gained this name because they crawl through a website browsing a page at a time, and then following the links to the other web pages on the website until every single one of the web pages have been read. This paper basically reviews the web crawlers: their architecture, their types and various issues and challenges being faced when the web search engines use these web crawlers.

KEYWORDS: Crawlers, types of web Crawlers, webpage, design issues.

I. INTRODUCTION

The World Wide Web is a collection of interlinked and hypertext documents which can be accessed via the Internet. Because the web is evolving at fast speed, user has to filter through several pages to come across the data one needs. In contrast to long established collections such as the libraries, the Web does not have a content structure that is centrally organized. This information can be downloaded with the help of web crawler. Search engines uses its components called crawler which helps the user to find the desired information. The basic job of a web crawler is to bring back web pages, parse them to get even more URLs, then bring these URLs to get back even more URLs. It is mainly a computer program that retrieves and accumulates pages in a storing repository.

Web crawling can also be seen as a graph search problem because as the web is considered to be a huge graph where nodes are web pages and hyperlinks can be called as edges. Web crawlers are be used in variety of areas, of which the most prominent is to index a large collection of pages and then allow people to search this created index. A Web crawler does not actually travel around computers that are connected to the Internet, as the viruses or intelligent agents do. It only sends the requests for data on web servers from a group of already known locations.

There are some important characteristics of the Web that make the process of crawling very difficult: large volume, fast rate of change, and the dynamic page generation. The large volume entails that the crawler can download only a fraction of the Web pages within a given amount of time, hence it needs to prioritize the various downloads. Secondly, the high rate of change mainly implies that by the time the crawler is downloading the last few pages from a given site, it is highly likely that new web pages have been added to the website, or some pages have been updated or deleted.

II. LITERATURE SURVEY

Web crawlers are used by many researchers to get the data from the web. Web crawling can be used to automatically discover and retrieve data from the web in web mining field. Pasquale De Meo, Salvatore A. Catanese, Emilio Ferrara [7], had used web crawlers on a social networking site i.e., FaceBook. The data accumulated by them was later used to learn the community structure of facebook. Priyanka-Saxena [8], introduced a new web crawler known as Mercator, which is a scalable web crawler that is written in java. Ari Pirkola [9], studied focused crawling to obtain biological data from the internet. Marc Najork and Christopher Olston [10] presented the fundamentals of web crawling. In his paper crawling architecture was discussed and also information about the future scope of crawling was given. Raja Iswary, Keshab Nath [11], discusses in his paper the various techniques to develop a web crawler and ways of building an efficient web crawler. Torsten Suel and Vladislav Shkapenyuk[12], discussed the design and the implementation of the web crawler. It is proposed by Vidal et. al. in his paper that searching of the desired web page(s)can be done more efficiently by recognizing the structure of a needed web page beforehand. Hence, instead of obtaining all web pages

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

related to the search topic, only those web pages which will have comparable structure as that of sample web page in terms of relevancy are fetched.

III. ARCHITECTURE OF WEB CRAWLER

One of the important components of the search engines is a web crawler. As the web is growing, there is growth in the use of web crawlers as well. Web crawler which are also called Spiders or robot, is program that starts with a Uniform Resource Locator (seed URL), then downloads the web pages with connected links and searches for the updates and then accumulates them for later use. This procedure is carried out iteratively[1]. The architecture of a crawler is given in Fig. 1.

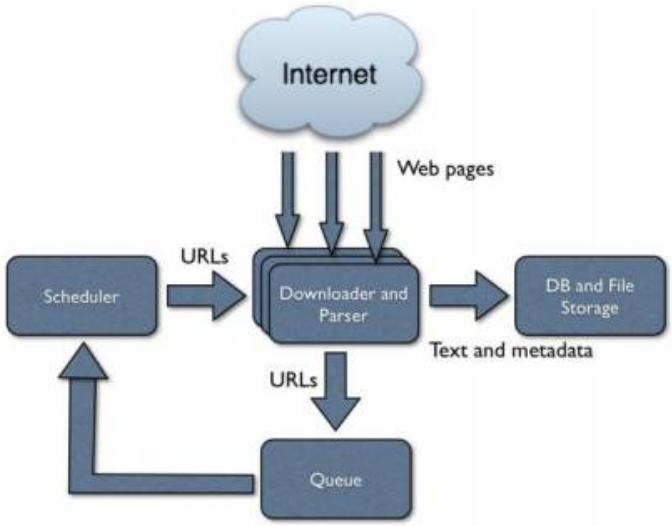

Fig 1. Architecture of web crawler

HTTP request is sent to the Web to download the web pages by mentioning the seed URL. The web pages are extracted by the web crawler and it follows the link given on that web page. It is then sent to the parser, which is a major component in the technology of crawling, which basically is used to check whether appropriate information is extracted. These relevant data are later indexed by the indexer and is then stored for use later. The web crawler searches for any updates on indexed web pages and if it is found, the old data is then replaced with the updated information, until the search term disappears from those web pages.

IV. TYPES OF CRAWLERS

A variety of crawlers with different strategies are available and are as follows:

A. Focused Web Crawler

This type of web crawler returns web pages which are specific and appropriate to the given topic. It decides how much the given web page is relevant to the given search topic and how to proceed further. Advantage of this web crawler is that it is feasible economically in terms of network resources and hardware. It employs variety of techniques for searching. Some focused crawlers also employ strategy of Best – Fit search. Whereas some focused crawler also employ page rank technique for finding out the most important web page [2].Variety of other techniques are also being used for focused crawling method.

B. Distributed Crawler

Many crawlers are distributed in the process of crawling the web so as to have the maximum coverage of the web. Central server supervises the nodes as they are geographically distributed and manages the communication and coordination among these distributed nodes. A large amount of computing and storage capability is needed which

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

increases the efficiency of crawling. Page Rank algorithms are mostly used for its increased efficiency of crawling. Other approaches are also proposed.

C. Incremental Crawler

This type of crawler is the one which on an incremental basis updates its collection of index once its accumulation of target is finally obtained. The existing collection is refreshed by performing new updates periodically. This is effective and also minimizes the use of network bandwidth. Several other approaches are also used.

D. Parallel Crawler

Parallel crawlers refer to the multiple web crawlers that often run parallel. These type of web crawlers depend on the web page freshness and the selection of the web page [2].This web crawler can be a parallel network or it can be distributed at very distant locations geographically.

V. WORKING OF A WEB CRAWLER

Crawling is the most delicate application because it involves interacting with many web servers. The speed of web crawling is not only governed by the speed of the internet, but is also governed by the speed of the web sites that are being crawled. Also the speed and time of crawling can be significantly reduced if there are a number of downloads happening in parallel.

A crawler roughly starts by placing in a queue, initial collection of URLs, where these URLs are kept and prioritized. Later, from this queue, the web crawler obtains the URL(in a given order), it downloads the web pages, retrieves the URLs(if there are any) from the downloaded web page and then puts these new URLs in the created queue. This procedure is repeated. The web crawler is used for crawling a whole website once a start-URL is specified and the crawler then follows every link found in that web page. A web site can be viewed as a tree structure, where the start URL is the root, and all links in that root web page are its direct children. Following links are then children of the previous children [4].

The basic algorithm of a crawler is [3]:

{
Start with seed page
{
Create a parse tree
Extract all URL's of the seed tree (front links)
Create a queue of extracted URL's
{
Fetch the URL"s from the queue and repeat
}
}
}Terminate with success

The various data structures that are involved in the web crawlers are as follows: -

Indexer: - A program that "reads" the web pages that are downloaded. Web indexing involves indexes to websites to provide a useful vocabulary for search engines. The HTML code is also examined by the indexer, which makes the web page and searches for terms that seem important. Different methods of indexing are available. Terms in italics, bold, header tags are given priority.

Repository: - It manages and stores a huge set of web pages. The web pages are stored in the repository by compressing. Different compression methods can be used for the compression pages, that is a tradeoff between compression ratio and speed. The documents are stored in a sequence in a repository.

Hit list: -It stores the list of words in the order of occurrence in a particular document which also includes position, capitalization, and font information. There are two different types of hits: plain hits and fancy hits.

Document index: - This keeps the data about every document.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

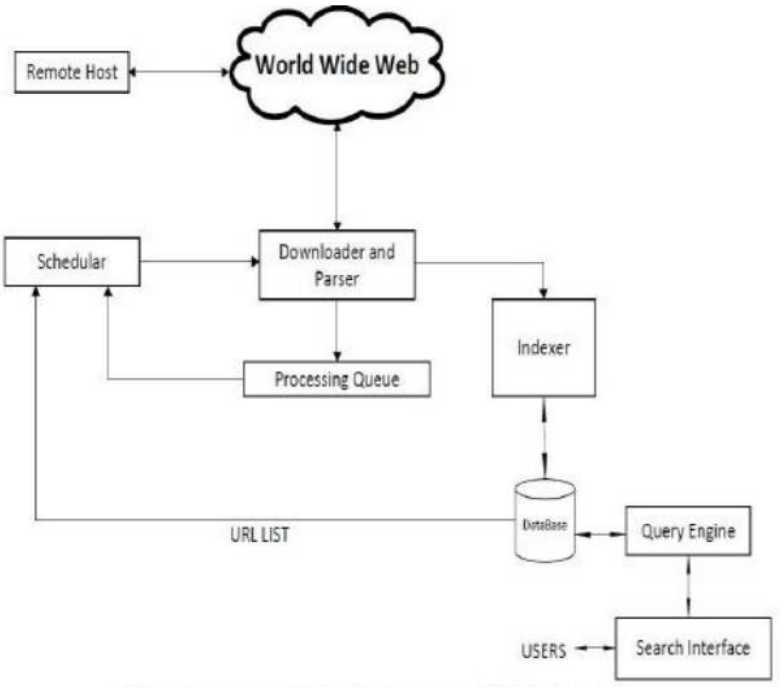

Fig 2 Working of Crawler

VI. APPLICATION OF CRAWLERS

There are a variety of uses of a web crawler: Crawlers can be used for automating the maintenance tasks on a given Web site, which includes validating the HTML codes or checking the links. Web crawlers can also be used to collect information of specific types from the Web pages, such as harvesting of an e-mail address (usually used in spam).
Linguists also use crawlers to carry out a textual analysis as they crawl the web to find out what terms are most commonly used these days.
Web crawlers are used frequently by search engines to collect data that is accessible on public web pages. This is done so that when the web surfers type a search term on their web site, they can quickly present the surfer with appropriate sites on the web.
Biomedical applications, for example: finding the related literature on a given gene, also uses the crawlers [1].
Market researchers need to determine and review trends in a market, for which they use web crawlers. Once the trends are studied, they can make good recommendation systems.

VII. CHALLENGES

i. *Cost*
For the owners of the web sites, the web crawlers may incur costs by using up the bandwidth allocation of the crawled websites. Many different web hosts are available that provide various server facilities, and also charging in many different ways. The penalty of surpassing the bandwidth results in higher costs to be paid so as to disable the website.

ii. *Scale*
The internet is growing day by day. Hence, for the web crawlers to achieve to wide coverage and excellent performance, it is needed to produce a high throughput. To rise above these problems, there is a need for the companies to employ a huge number of computers which will count to thousands and also dozens of high speed links of network.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

iii. *Privacy*
The privacy issue for crawlers appears to be clear-cut since everything on the internet is in the public domain. Information o the web may still lead to invasion of privacy if it is approached in certain ways, especially when data is combined on a huge scale over many web pages [1].

iv. *Social Obligations*
In order to avoid the denial-of-service kind of attacks, web crawlers should follow safety means. The web crawlers should set up a good quality coordination with the different sites for which perform the work.

VIII. CONCLUSION

This paper describes the web crawlers and its type and also discusses the architecture and the working of the web crawler in general. A variety of applications of the crawlers are also discussed. We also saw the various challenges of a web crawler. The web crawlers are an important part of all the search engines. Web crawlers need to provide high performance as they are the basic components of the web services. Constructing an effective crawler to solve purposes is not a difficult job, but the choosing of the accurate strategies and constructing an effective architecture will surely lead to implementation of intelligent applications of the web crawler.

REFERENCES

[1] Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica "Web Crawler: Extracting the Web Data", International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014.
[2] Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli "Web Crawler in Mobile Systems", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.
[3] Akshaya Kubba "Web Crawlers for Semantic Web", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015.
[4] Shalini Sharma "Web Crawler", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014.
[5] Deepika, Dr Ashutosh Dixit "Web Crawler Design Issues: A Review", IJMIE, Volume 2, Issue 8, August 2012.
[6] Subhendu kumar pani, Deepak Mohapatra, Bikram Keshari Ratha "Integration of Web mining and web crawler: Relevance and State of Art", International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010.
[7] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, Alessandro Provetti, "Crawling Facebook for Social Network Analysis Purposes", WIMS'11, May 25-27, 2011 Sogndal, Norway, 2011.
[8] Priyanka-Saxena "Mercator as a web crawler", International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, ISSN: 1694-0814, 2012.
[9] Ari Pirkola, "Focused Crawling: A Means to Acquire Biological Data from the Web", University of Tampere Finland, ACM 978-1-59593- 649-3/07/09, 2007.
[10] Christopher Olston and Marc Najork, "Web Crawling", now the essence of knowledge, Vol. 4, No. 3,2010.
[11] Raja Iswary, Keshab Nath "Web Crawler", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 10, ISSN: 2278-1021,2013.
[12] Vladislav Shkapenyuk, Torsten Suel, "Design and Implementation of a High-Performance Distributed Web Crawler", NSF CAREER Award, CCR-0093400.