# DESIGN AND IMPLEMENTATION OF WEB CRAWLER

[1]Rucha Chute, [2]Sai Bawiskar, [3]Khushboo Mishra,[4]Jeet Singh Paleriya
Department Of Information Technology,RGCER, Nagpur.
Email:[1]ruchachute94@gmail.com, [2]sai22kar@gmail.com, [3]khushboomishra559@gmail.com,
[4]jeetsinghpaleriya@gmail.com

**ABSTRACT:**

**The large size and the dynamic nature of the Web highlight the need for continuous support and updating of Web based retrieving information system. Web crawlers are programs that exploit the graph structure of the Web to move from page to page. When an user initiates a search, the key words are extracted and searches the index for the websites which are most relevant. The size of the web is huge, search engines practically can't be able to cover all the websites. There should be high chances of the relevant pages in the first few downloads,as the web crawler always download web pages (in fractions). Implementing the proposed method, we download web page and get the title, body and the number of outgoing links on that particular page in order to calculate relevancy.**

**Keyword :**

Web crawler: A web crawler is a program that, given one or more seed URLs,downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks.

Seed: It is starting URL from where Web Crawler starts traversing World Wide Web.

Frontier: It is list of unvisited URLs.

Page weight: Weight of page which is decided on the certain parameters.

Threshold value: Certain limit which we decide the importance of page.

## INTRODUCTION

Internet is the shared global computing network. It enables global communications between all connected computing devices. It provides the platform for web services and the World Wide Web. Web is the totality of web pages stored on web servers. There is a spectacular growth in web-based information sources and services. It is estimated that, there is approximately doubling of web pages each year. As the Web grows grander and more diverse, search engines also have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated. In Internet data are highly unstructured which makes it extremely difficult to search and retrieve valuable information. Search engines define content by keywords. With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources

A web crawler(also known as a *robot* or a *spider*) is a system for the bulk downloading of web

pages. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries.

Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. Web crawlers are programs that exploit the graph structure of the Web to move from page to page. In their infancy such programs were also called wanderers, robots, spiders and worms, words that are quite evocative of Web imagery. . The large size and the dynamic nature of the Web highlight the need for continuous support and updating of Web based information retrieval systems The last key dimension is regarding crawler evaluation strategies necessary to make comparisons and determine circumstances under which one or the other crawlers work best. Crawlers facilitate the process by following the hyperlinks in Web pages to automatically download a partial snapshot of the Web.

**Features a crawler should provide**

**Distributed:**The crawler should have the ability to execute in a distributedfashion across multiple machines.

**Scalable:**The crawler architecture should permit scaling up the crawl rate by adding extra machines and bandwidth. Performance and efficiency: The crawl system should make efficient use of various system resources including processor, storage and network bandwidth.

**Quality:**Given that a significant fraction of all web pages are of poor utilityfor serving user query needs, the crawler should be biased towards fetching "useful" pages first.

**Freshness:**In many applications, the crawler should operate in continuous mode: it should obtain fresh copies of previously fetched pages. A search engine crawler, for instance, can thus ensure that the search engine's index contains a fairly current representation of each indexed web page. For such continuous crawling, a crawler should be able to crawl a page with a frequency that approximates the rate of change of that page.

**Extensible**: Crawlers should be designed to be extensible in many ways –to cope with new data formats, new fetch protocols, and so on. This demands that the crawler architecture be modular.

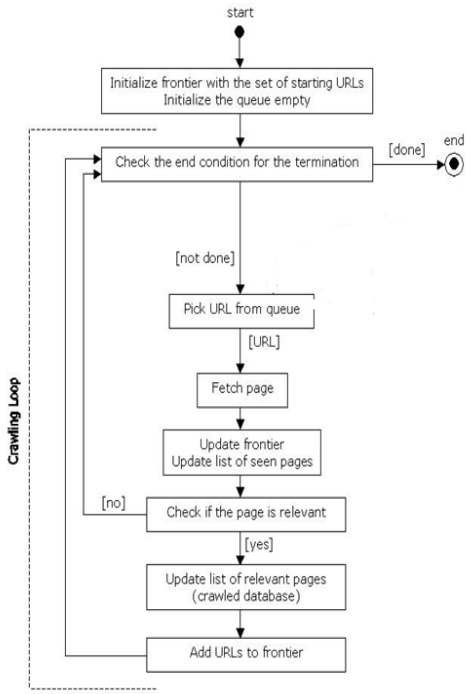



***Fig : Flowchart of Web Crawler***

I. Materials & Methods

We define the factors for which we specify the page importance:

weight(page) = weight(URL) + weight(outlinks) + weight(title) + weight(body) where,

1) if ( search string present in URL)
   {
   weight(URL) returns a predefined weight
   }
   Else
   {
   Return 0
   }
   This will return the weight assigned for the URL occurrence. If the search string is found in the URL, the page acquires certain importance.
2) if ( search string present in title)
   {
   weight(title) returns a predefined weight
   }
   Else
   {
   Return 0
   }

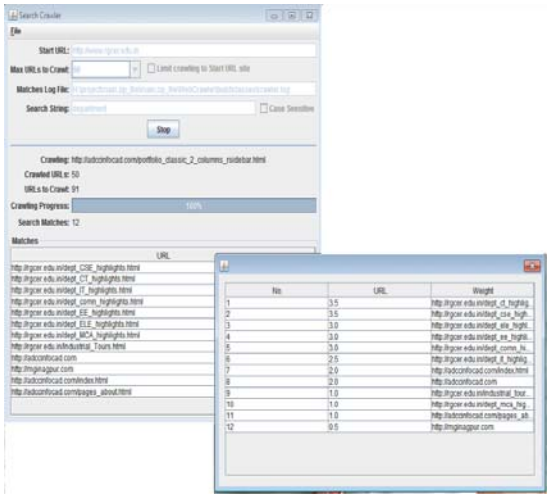
This will return the weight assigned for the title occurrence. If the search string is found in the title, the page acquires certain importance.

3) Occurrence of search string in the body
{
weight(body)=occurrence*weight for each occurence
}
This will return the weight assigned for the body occurrence. If the search string is found in the body, the page acquires certain importance. When the search string occurs certain number of times in the body, the occurrence is noted and the page importance is calculated using the occurrence count.

4) Number of hyperlinks on the page
{
weight(outlinks)=occurrence*weight for each occurence
}
This will return the weight assigned for the out-links occurrence. The number of links linking to the other page has also been assigned some importance. Giving importance to each component of the parsed page, we have assigned weight to each component and hence acquired the page importance in totality. As we get the page weight, we will compare it with the threshold frequency implicitly provided to the algorithm. Depending on the result of comparison, the links are either added to the output or they may be discarded. Thus, we get the search more focused to the search string eliminating the least.

## II. RESULT & DISCUSSION

Seed URL : http://www.rgcer.edu.in
Search String : Department



Seed URL passed is http://www.rgcer.edu.in and the search string was department. Matched url i.e. Relevant url out of total 50 urls were 12 . Crawler had calculated 24% relevant url's out of total URLs.

Calculations of Result 1

$$\frac{\text{Relevant URLs}}{\text{Total URLs}} = \frac{12}{50} * 100 = 24\%$$

## III. CONCLUSION

Hence by using the concept of Page Weight, we scan web pages as well as compute the weight of page and hence we can increase efficiency of web crawler as output set of URL generated by this way will always be of better importance than what traditional web crawler is generating. Hence we have obtained the pages with their respective weights and compared them with the threshold weight. Hence acquiring the more relevant pages.

## REFERENCE


1. From Wikipedia -"Web Crawler".
2. P. Gupta & K. Johari "Implementation Of Web Crawler",Emerging Trends In Engineering & Technology (ICETET) Second International Conference,Nagpur,India,2009.
3. S.MALI, "Implementation of multiuser personal web crawler",Software Engineering (CONSEG) CSI Sixth InternationalConference,Nagpur,India,2012.
4. www.devarticles.com.