# Effective Searching Policies for Web Crawler

Suyash Gupta, KrishanDev Mishra, Prerna Singh
(Department of Information Technology, Mahamaya Technical University, India)
(Department of Information Technology, Mahamaya Technical University, India)
(Department of Information Technology, Mahamaya Technical University, India)

***ABSTRACT:*** *As we know search engines cannot index every Web page, due to limited storage, bandwidth, computational resources and the dynamic nature of the web. It cannot monitored continuously all parts of the web for changes. Therefore it is important to develop effective searching policies. In this technique there is the combination of different searching technique to form a effective searching policies. These combined techniques are best first search, focused crawling, info spiders, recrawling pages for updation, crawling the hidden web page.This combined technique also includes Selection policy such as page rank, path ascending, focused crawling Revisit policy such as freshness , age Politeness , Parallelization so that it allow distributed web crawling.*

***Keywords:*** *Best first search, recrawling pages for updation, crawling the hidden web page, focused crawling, Revisit policy, Politeness, Parallelization.*

## I. INTRODUCTION

A Web crawler is a software program that helps in locating information stored on a computer system, typically based on WWW. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses.

Given an initial set of seed URLs[5], it recursively downloads every page that is linked from pages in the set. A focused web crawler[10] downloads only those pages whose content satisfies some criterion also known as a web spider, bot, harvester.Web crawlersare commonly known as a Web Spider or a Web Bot. The crawler is an automated program that is designed to scan the World Wide Web. These crawlers are used by search engines like Google, Yahoo or MSN.

It provided these engines with update knowledge of the sites that are on the web. The crawlers begin with a list of sites to scan, these are called seeds. When the crawler comes to the sites it first identifies the hyperlinks on the page and then puts them on another list of the Url's called the crawl frontier. There are a couple of factors that may affect the way a crawler crawls a site. If there is a large volume of information it may take the crawler a bit of time to send the downloads. So the crawler will have to see what information it wants to send down first.

Another thing a crawler will notice is if the site has undergone a change since the last time it has crawled that site. There may have been pages added or removed during this time, which is a good thing because a search engine want to see a fresh site as long as you remove old pages. It is single piece software with two different functions building indexes of web pages and navigate the web automatically on demand.

Our goal is to have the crawler for our site in order to get a good rank. It can take a crawler weeks or a few months to change a rank but in that time there could be changes that have been made to the site. The crawler will already have taken this into consideration. You may have added or removed some content. Try to keep the site the same until your index changes. You can add pages because search engines love new content.They help us in getting track or getting updated information which we want to access.

## II. HISTORY OF WEB CRAWLER

WebCrawler was the first web search engine. It was used to provide full text search. It was bought by America Online on June 1, 1995 and sold to Excite on April 1, 1997. Excite was born in February 1993 as a university project called Architext seeking to use statistical analysis of word relation to improve relevancy of searches on the internet. Then, web crawler was acquired by InfoSpace in 2001 after Excite, went bankrupt. InfoSpace also made to operates the metasearch engines DogPile and MetaCrawler. More recently it has been repositioned as a metasearch engine, providing a composite of separately identified sponsored and non-sponsored search results from most of the popular search engines.WebCrawler also changed its image in early 2008, scrapping its classic spider mascot.In July 2010, WebCrawler was ranked the 753rd most popular website in the U.S., and 2994th most popular in the world by Alexa. Quantcast estimated 1.7 million unique U.S. visitors a month, while Compete estimated 7,015,395 -- a difference so large that at least one of the companies has faulty methods, according to Alexa.[2][3][4]

## III. WORKING OF WEB CRAWLER

Web crawler collects all the documents from the web to build a searchable index for the search engine. After collection and build a searchable index the web crawler identifies all the hyperlinks in the pages and adds them to the list of URLs to visit. The web crawler continuously populates the index server. The process of the search engines queries are firstly, the web server sends the query to the index servers that tell which pages contain the words that match the query.

Secondly, the query travels to the doc server, which actually retrieve the stored documents, snippets are generated to describe each search result. Thirdly, the search results are returned to the user. It recursively downloads every page that is linked from pages in the set. Our main goal is that the crawler comes to our site in order to get a good rank (i.e. highly ranked) for you. It can be done by having something on our site that a crawler will not see kind of defeats the purpose of having a site. It can take a crawler weeks or a few months to change a rank but in that time there could be changes that have been made to the site. In this there are combination of different technique for parameter tuning so that it shows the results of the searches as per the highly rank. It will also help to remove unwanted pages on the web[6].

## IV. CATEGORIES OF SEARCH ENGINE : in terms of how they work

**Crawler based search engine:** These search engines create their listings automatically. Examples are Google, Yahoo. The Google web crawler will enter your domain and scan every page of your website, extracting page titles, descriptions, keywords, and links – then report back to Google IQ and add the information to their huge database. They crawl or spider the Web to create directory of information. People can search through the directory created in the above process. When changes are made to page, such search engines will find these changes eventually because they are automatically go from one web server to other looking for information, based on whatever information it can find their it will try to build up a directory. Later on whenever user submit a query the directory will be consulted to return the result. And if some pages are updated then the next time when the crawler again visit the web the new updated version will be consulted on the directory. It can serve automatically[8].

**Human powered search engine:**These depend on humans for the creation of directory. Example: OpenDirectory. One submits a short description (contain keyword and other information) for the web site to be listed to the directory. When searching is response to some user queries, only the description submitted are searched for. This process has been carried out by the person who had created the web page.Alternatively, editors can write reviews for some web sites. The editors will be taken a responsibility of submitting a information to the directory service, so that the web page is to be indexed. When changes are made to the page, it has no effect on the listing[8].

**Hybrid search Engine:** It can accept both types of result based on web crawlers andbased on human-powered listings. Hybrid search engine can look through dictionaries which are created by web crawlers and also using human power listing so both these are allowed there. But however most of such hybrid engines can assign prioritiesout of the web crawler and human powered listing which one should given higher priority. For example MSN search (product of Microsoft) gives priority to human powered listings and the technology of the tool that is LookSmart.MSN search also presents crawler based search results for complex queries. So it first look for the human powered listings and then for other one. The tools that it uses for crawler based search are Inktomi, SubmitExpress[8].

## V. COMPONENTS OF CRAWLER BASED SEARCH ENGINES

Broadly they have three major elements first one is**crawler or spider**[1] that implies crawler will crawl from one web site to other. Usually this crawling is based on the hyperlinks that are present on one web page. It visits a web page, retrieves it, and follows the hyperlinks to other pages within the site. It visit the site regularly(once in every month) and look for changes that had taken place on the pages since the last time it was visited. Second one is **Index or catalog**[7]. It is like a huge book containing a copy of every web page that the crawler finds. It means the crawler whatever pages it retrieves it get stored in the index or catalog. Once stored it will remain there and it will be updated when a page changes. Indexing is an important step until a page is indexedit is not available for search. The Third one is **search engines software**is the actual softwarethrough which the user submit the queries. This program searches through the million of entries in the index because the index will typically contain huge number of entries to find the matches to a search. It can also rank the matches based on relevance (they should be mechanism for us to tell or to find out whether the particular page is relevant for search or not or given alternate pages). All crawler-based engines have above basic components, but differ in the ways these are implemented are tuned.

## VI. EVALUATION OF CRAWLER

In a general,a crawler may be evaluated on its ability to retrieve good pages. However, a major hurdle is the problem of recognizing these good pages. The real users may judge the relevance of the pages as these are allowing the users to determine if the crawled was successful or not in an operational environment.

As we know the evaluation of crawler is quiet difficult because the main purpose of the crawler is to retrieve good pages, in the sense that are beneficial to the user for which they looked for. It is very important to judge the relevance of the pages that are crawled by crawler. Search engines cannot index every WEB page due to limited bandwidth, storage and computational resources and to the dynamic nature of the WEB, therefore it is important to develop effective crawling methods to assign priorities of the pages to be indexed. Since here we are not involving any real users, here we use topics instead of queries, each represented by a collection of seed URLs (uniform resource locator). Now it is clear that we clear the issues by moving from queries to topics. Let us consider we may lose any clues to user context and goals that queries may provide. However, this approach of beginning with the seed URLs is increasingly common in crawler research. We assume that if a page is on topic then it is a good page. There are also some limitations with assumption. All the assumption may not be sufficient condition according to user relevance. For example, a user who has already viewed a topical page may not consider it relevant since it lacks novelty. While we don’t under rate these criteria, given the reasons stated above we choose to focus only on topic as an indicator of relevance for the extent of this research.[9][10]

## VII. CONCLUSIONS

In past there is not very well documented in the written that every web crawler is scalable and also everything in the world has some limitation. These limitations are like parameter tuning, remove unwanted pages, paid crawl ordering, deep web. So in order to remove this limitation we have to make the effective searching policies for web crawler get accurate result. In this method we combine different technique which results into one can easily get the appropriateresult, it will also help to remove the unwanted web pages because of finite storage capacity.

## REFERENCES

[1] He, H., Meng, W., Yu, C., and Wu, Z. WISE-Integrator*: A System for Extracting and Integrating Complex Web Search Interfaces on the Deep Web*. In Proceedings of the 31th VLDB Conference. 2005.

[2] Webcrawler.com Site Info Alexa, 2010

[3] Webcrawler.com-- Quantcast Audience Profile", Quantcast, 2010

[4] Site Profile for webcrawler.com, Compete, 2010

[5] A. Broder, M. Najork, and J. Wiener, "*Efficient URL caching for World Wide Web crawling,*" in Proceedings of the 12th International World Wide Web Conference, 2003.

[6] Grossan, B. "*Search Engines: What they are, how they work, and practical suggestions for getting the most out of them*," February1997. http://www.webreference.com.

[7] C. Duda, G. Frey, D. Kossmann, and C. Zhou, "*AJAXSearch: Crawling, indexing and searching web 2.0 applications*," in Proceedings of the 34th International Conference on Very Large Data Bases, 2008.

[8] I.Sengupta, "*Search Engine And Web Crawler*", Dept of computer science and engineering, Indian Institute of Technology Kharagpur.

[9] FilippoMenczer, Gautam Pant, PadminiSrinivasan, Miguel E. Ruiz," *Evaluating TopicDriven Web Crawlers*", Department of Management Sciences, School of Library and Information Science, The University of Iowa, Iowa City, 2001

[10] Y. Yang. "*An evaluation of statistical approaches to text categorization*", Information Retrieval, 1(1):69-90,1999.