IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12808

Effective Information Retrieval using Web Crawlers

Dr. Kompal Aggarwal

Assistant Professor, Govt. College, Panchkula

Abstract: Most people prefer search engines as their initial method of browsing information on the internet. Most of the time, the returned information is not the anticipated search material or is not requires. Search engines progress continuously for fulling human search behaviour for returning the best information available and to identify the search query. Search engines have a time to decide which is the most proper information to give to the searcher since, in these days, the information is extensive and so numerous. The Web crawlers predicts the searching quality on behalf of search engines.

I. BACKGROUND OF THE STUDY

Search engines [1] offer the people who browse with the information they want with a unique reason.. The technique utilized by search engines predicts the searching quality. The techniques must give the nearest results in increasing order and guess the intended information requested by the user so that the nearest result will appear first in a SERP [2]. The search technique, which is used for the accuracy of the search results displayed on the SERP, is completely dependent on the user. The user exactly does not know how the search engine decides what results are to be displayed since the techniques are closely guarded secrets by the search engines. Since much amount of information is unknown about the techniques being used in search engines, for predicting how much it affects the search results by analyzing the results given by the search engines and viewing at the search outcomes, the research uses one of the search methods known as keywords. The most popular search engines used nowadays are the “Yahoo, Bing and Google”. Since most people tried to check out the work of search engines, the SEO organizations notice their clients craft their client’s websites come into view at the top of SERPs. The real intention is that no one wants their hard work to be left unknown by netizens and everyone needs their document or website to be easily searched and visible. The positive side of SEO relies on the forefront of most marketing campaigns and the negative side of SEO relies on the forefront of marketers and hackers of uncertain products. Since search engines are conscious about this fact, it has made the globe of search engines even more complex because search engines exclude documents that have uncertain content and made it to the top of SERPs.

II. HISTORY OF SEARCH ENGINE

`Since the first search engine, Archie was discovered in 1990, it has come a long way. Before Google underestimated the whole search arena, many search engines were discovered after 1990 that included “AltaVista, Yahoo, InfoSeek and Excite”. Google [3] was invented in 1996 by “Sergey Brin and Larry Page”. Since other search engines drifted into insignificance, Google remained much powerful since its beginning. One of the major reasons Google earned popularity and distinguished itself from the rest is because of its simple design while the rest messed their search page with advertisements and images. Another major aspect was Google’s creative performance of the PageRank factor that ranked websites relative to their consequence. When numerous SEOs [4] started to multiply in masses to take gain of Google’s PageRank technology, certain websites make them visible at the top of Google’s search listing by jerking them. PageRank has seen its significance being played down by Google as its fame also proved to be its vengeance, thereby making Google’s search results seem manipulated.

For determining whether the results of a search are related to a query, a computer must decide on the following. A document will be deemed more pertinent if the following conditions occur:

a) If the document is involved with numerous query terms
b) If there is a possibility of occurring the query terms more frequently
c) If the document is involved with less non-query terms

A good indexing system will be capable to reconstruct new indexes, do not need large resources and answer queries effectively and quickly, all of which are simply done with the accessibility of present computing power. The hard part

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.12808

is the consequence of the search result and shaping the effectiveness of the answer. The most familiar way to determine the consequence of a search result is to compute how early the documents occur in the ranked list and how many of them have been retrieved. The three quantities that calculate the retrieval performance is as follows:

a) The count of retrieved documents
b) The count of useful retrieved documents
c) The total count of applicable documents

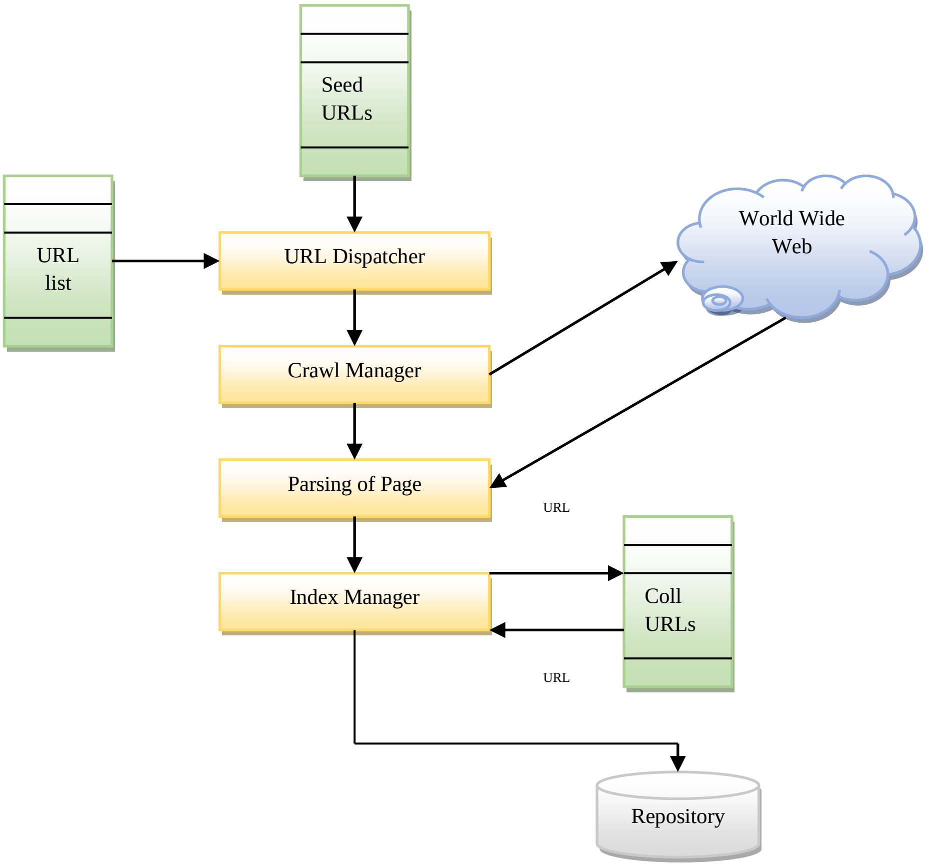

Figure 1.1: Architecture of a Web Search Engine

The query engine fills and receives the request from the user while searching. It depends on the repository and the indexes. Result sets are generally very big because of the size of the web and the data that is based on the entered keyword. SEO can be utilized to enlarge the number of desired visitors to a website through search engines and to illustrate a diverse set of activities [5]. These include actions such as pursuing other sources of traffic for links or listings, communicating directly with the search engines and making changes to the text and HTML code.

SEO positions well for particular keywords that visitors will explore for in search engines and it is fundamentally doing things to set up a website. SEO initialized via commentaries and public reports given by search engine experts. Early information about SEO viewed at how the different search engines ranked results of search and search engine techniques. Website owners and inspired entrepreneurs started learning these tested strategies and reports on how they can rank better on the search outcomes.

As expected, it soon became a profession and a business for some to be occupied in the services for supporting others and providing SEO advice to support them in attaining better rankings for their websites. As WWW grew at a significant pace, Google became bigger and stronger and started deteriorating the fame of some search engines like Infoseek and Alta Vista. Google's early success was due to the statement that their web interface was less clustered with advertisements and their groundbreaking new technique which made use of their proprietary PageRank formula.

Google today has developed to be the well-performing search engine holding over 60 percent in market share and is much more difficult. SEO companies understand that everyone needs to obtain their website listed on Google to draw as much traffic as possible from organic search and this has invariably made them the target for almost every SEO company. Nowadays many companies do not have the time to certify that their website will be well ranked as they may not be well-known as to what the various search engines are looking for and there are also many who practiced SEO unprofessionally to attain traffic to their websites and it cannot be deprived that SEO is a much wanted after service.

Despite thinking about whether their design or content will be liked by the search engines, many website creators are logically more anxious about the design and content they place inside their webpages. SEO is significant since very few developers develop websites where visitors or traffic are not required and it cannot be denied by a website developer [6]. The first few pages of the search results are barely looked at by the search engine visitors. Since there are probably thousands of other similar websites, there is a need to take the risk being sidelined by the search engines or compensate heed to what SEO is all about. Google does not pay to understand how it works if it is in the traffic since it is the preferred search engine of the day.

III. CRAWLER-BASED SEARCH ENGINE

With the support of web crawlers, crawler-based search engines produce their listings routinely. A computer technique is utilized to rank all retrieved pages. Such search engines frequently retrieve a lot of information from the web and they are vast. It enables us to filter search results and allows exploring within the results of the previous search in terms of complex searches. The full text of web pages in which they linked is enclosed in such types of search engines[7]. Pages can be discovered by matching words in the pages that the user needs . Some examples of crawler-based search engines include “Altavista, Lycos, Teoma, AllTheWeb and Google etc”.

There are two fundamental takes that the crawler should follow such as keeping the freshness of previously downloaded pages and downloading the new pages. Good freshness can only be assured by placing the unwanted burden on the crawler and by merely revisiting all the pages regularly which is not feasible. It is necessary to crawl the web in a scalable and competent way if the reasonable measure of freshness or quality is to be managed with the accessible bandwidth for performing crawls that is neither free nor infinite. A crawler must possess features like scalability, robustness etc. Web crawlers are utilized by many sites and web search engines to modify their indexes or web content of other sites. A web crawler is a robotic crawler or an Internet bot that analytically browses the WWW, usually for the intention of web indexing that supports faster access of information. A web crawler has many names like “Robot, Wanderer, Robot, worm, Web agent etc”. The globally faced major issues in web crawling are:

- Execution time
- Collaborative web crawling
- Scale: Millions of pages on web
- Crawling the large web repository
- No central control of web pages
- Crawling multimedia data
- Large number of webpages emerging daily

The web crawlers track links to contact different pages and they are perhaps small programs that scrutinize the search engine’s web[8]. Crawlers accumulate pages in a repository database and capture the URLs showing the retrieved pages. Based on the host protocol, it chooses a URL from its seed URLs and downloads the documents from the web server. The browser makes it available to the user and parses the document.

The crawler initiates by locating an initial set of seed URLs in a queue, in which the entire URLs to be retrieved are prioritized and kept. The crawler downloads the page from the corresponding queue and places the new URLs in the queue, extracts a URL and extracts URLs from the downloaded page.

The gathered pages are utilized by a search engine and the technique is repeated. The browser makes the document available to the users and parses them. The common architecture of a focused crawler is shown in Fig 1.2.

DOI: 10.17148/IJARCCE.2023.12808

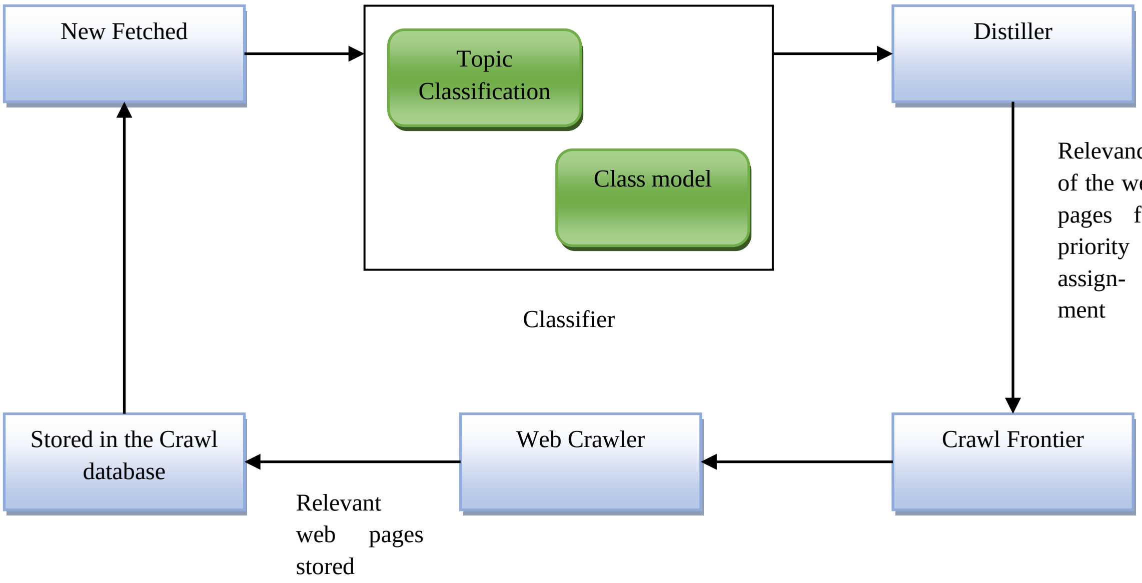
New Fetched
Topic Classification
Class model
Classifier
Distiller
Relevancy of the web pages for priority assign-ment
Crawl Frontier
Web Crawler
Stored in the Crawl database
Relevant web pages stored

Figure 1.2: Broad Architecture of a Focused Crawler

Maintaining the possibility of lowering the average age of pages and the possibility of enhancing the freshness of pages is the key role of a focused web crawler. i.e. the crawler deal with checking the old copies as well as updated copies. Downloading all pages of the web by the crawler is not possible in most of the cases. A little fraction of the whole web is presently indexed by the most widespread search engine. Another significant task of a crawler is selecting the important pages cautiously and visiting them first so that the information gathered in the local repository is more useful and the helpful visited web. The crawler consumes the other organization's resources when gathering pages from the web. There is a requirement for using the file system to retrieve page, consumes CPU and disk resources at once the crawler downloads pages from a site. The impact on the resources must be lessened by the crawler. Sometimes, the network may entirely block to be accessed by the crawler and the administrator of a particular network or website may complain. Crawlers download pages in parallel and often run on multiple machines owing to the huge web size . For downloading the numerous pages in a given time, this parallelization is frequently needed. Various crawlers do not visit a similar website frequently if these parallel crawlers are synchronized accurately. However, this coordination restricts the number of simultaneous crawlers, thereby incurring considerable communication overhead.

REFERENCES

[1]. A. Rungsawang and N. Angkawattanawit, "Learnable topic-specific web crawler", Journal of Network and Computer Applications, vol. 28, no. 2, pp. 97-114, April 2005.

[2]. Su Guiyang, Li Jianhua, Ma Yinghua, Li Shenghong and Song Juping, "New focused crawling algorithm", Journal of Systems Engineering and Electronics, vol. 16, no. 1, pp. 199-203, 2005.

[3]. A. Rungsawang and N. Angkawattanawit, "Learnable topic-specific web crawler", Journal of Network and Computer Applications, vol. 28, no. 2, pp. 97-114, April 2005.

[4]. Hai Dong and Farookh Khadeer Hussain, "Focused Crawling for Automatic Service Discovery, Annotation, and Classification in Industrial Digital Ecosystems", IEEE Transactions on Industrial Electronics, vol. 58, no. 6, pp. 2106-2116, 2011.

[5]. Manish Kumar, Ankit Bindal, Robin Gautam, Rajesh Bhatia, "Keyword query based focused Web crawler", Procedia Computer Science, vol. 125, pp. 584-590, 2018.

[6]. Gunjan H. Agre and Nikita V.Mahaja, "Keyword focused Web Crawler", 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, IEEE, pp. 1089-1092, Feb 2015.

[7]. Gunjan Agre and Snehlata Dongre, "A Keyword Focused Web Crawler Using Domain Engineering and Ontology", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue3, pp. 463-465, Mar 2015.

[8]. F. M. Javed Mehedi, Shamrat, Zarrin Tasnim, A.K.M Sazzadur Rahman, Naimul Islam Nobel, Syed Akhter Hossain "An Effective Implementation Of Web Crawling Technology To Retrieve Data From The World Wide Web (WWW)" International Journal Of Scientific & Technology Research, Volume 9, Issue 01, January 2020 ISSN 2277-8616