

Search Engine Optimization Using Crawler Optimization- A Technique

Rajesh Singh¹, S.K. Gupta²

¹Scholar, Bhagwant University, Ajmer, Raj.

²Proff, BIET, Jhansi, U.P.

Abstract:- The use of web search engine is an essential part of daily life. It is typical and impossible to underestimate the tremendous role they have for internet users. With today's information knowledge in enormous forms, as well as N/W information into the exponential grows of the trend, the search engine are more essential in our life. Generally Search engine provide search results without consideration of interest or content of user. We propose a new search engine architecture with robots in it.

Keywords:- SEO, Robots, Detection engine, discarded pages.

1. Introduction:-

Since the web based search engine appeared in 1994, the search engine has achieved great development. It aims for information navigation offering the retrieval service for users by searching, gathering and understanding, extracting, organizing and processing the information on internet with certain tactics. Traditional search engine comprises of three core components, download, index and search. A copy of all visited pages is stored for later use, Normal search engine returns thousands of matches in response to a user query. The size of information is too large to go through.

Existing search engine suffer from following drawbacks-

- Crawler are not able to analyze the context of the keyword in the web pages before they download it.
- The user submit his requests for retrieval of information without explicitly mentioning the context in which he or she desire.
- Crawler treats user search requests in isolation.
- Results returned are identical, independent of the interest of user.
- There is a need to prepare separate files for each web document.
- Augmentation is required in HTML document.

2. Literature Review:-

Search engine optimization is the process improving the visibility of a website or webpage in search engine SEO may target different kind of search like image search, local search, video search, academic search, news search, industry specific vertical search. Search engine is a term

used for information retrieval. It match queries against an index that they create. This index contain the words in each document, pointers to their location within the document.

Search engine has four essential modules:-

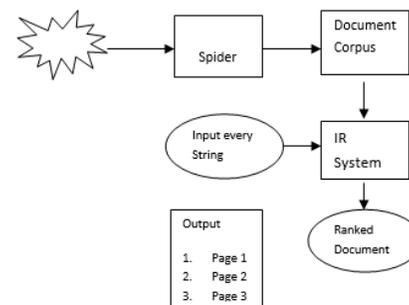
1. Document processor
2. Query Processor
3. Search and Matching function
4. Ranking Capability

Example of Search Engine are Alta Vista, Google, Hot Bot, MSN.

In other word we can define the Parts of any search engine are

- (1)Crawler or Spider
- (2)Index
- (3)Search engine Software

General Architecture of Search Engine is as follows



Each and every Search Engine should be able to produce the most relevant results to any given search. A true search engine is an automated software program that moves around the web collecting WebPages to include in its catalogue on database web crawler is a program that retrieves and stores pages from the web. Web crawler starts with a list of URL's to visit called the seeds. As the crawler visit these URL's it identifies all the hyperlinks in the page and adds them to the list of URL's to visit called the crawl frontier. URL's from the frontier are recursively visited according to asset of policies. Web Crawler are the building block of search engines. There is a URL server that sends lists of URL's to be fetched to the crawlers. The WebPages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into repository. Every web page has an

associated ID number called a doc-ID which is assigned whenever a new URL's is passed out of a web page. The indexes and sorter perform the indexing function. Indexer reads the repository, uncompressed the documents and passes them.

3. Need Of Study:-

There have been several studies of web crawling in its relatively short history, but most of them have had a focus rather different from ours. Some have concentrated on aspects relating to caching and. Others have been principally interested in the most accurate and effective way to update axed size database extracted from the web, often for some specific function, such as data mining. Several authors, e.g., Co man approach crawling from a theoretical point of view, comparing it to the polling systems of queuing theory, i.e., multiple queue-single server systems.. A common assumption has been that page changes are a Poisson or memory less process, with parameter as the rate of change for the pages. Nevertheless, the widely accepted Poisson model forms the basis for a series of studies on crawler strategies. These lead to a variety of analytical models designed to minimize the age or maximize the freshness of a collection by investigating:

how often a page should be crawled. In what order pages should be crawled Should a crawling strategy be based on the importance Of pages or their rates of change The average size of individual pages is growing the proportion of visual and other nontextual material is growing in comparison to text The number of pages has been growing exponentially low . Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available on the World Wide Web. To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from last years

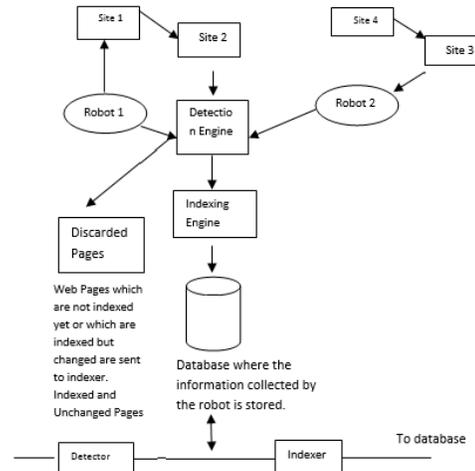
4. Proposed System:-

This architecture exploits the techniques of web Page change detection followed by indexing of the pages which are not yet indexed or the pages in which the content has been modified.

In this architecture Search engine is as a queuing system which has two steps to the web pages extracted by the web crawler.

Function of such type of Architecture is as follows.

- 1) Web Pages extracted by the web crawlers are queued up for the detection of the Search engine which has stage 1 service of the queue model.
- 2) Web Pages are then serviced by Detector in it the message digest of web page is computed and then matched with database of indexed web pages.
- 3) In the case of 18 Page has been indexed by search engine and respective message of the page is same as that stored in database then there is no need to index the page, but in case Page has not been indexed or indexed then the Page need to be indexed or re indexed, because message digest of Page changes only when the content of page has been modified.
- 4) Page corresponding to previous condition are discarded by the detector.



Algorithm:-

- (1) Input page to detection engine.
- (2) For each page in detector input queue, do:
 - (a) If page has not been indexed yet, send the page to indexing engine queue.
 - (b) If page has been already indexed, but changed, send the page to the indexing engine queue.
 - (c) If page has been already indexed and unmodified discard the page
- (3) Exit

5. Conclusion:-

The Personalized Search provide a common interface and conducts searches in many search engines simultaneously and provide results an proper manner- In present days search engine are useful tools to extract the required information from internet. This Architecture solve this purpose with big span of coverage and advanced features like using two different Robots, Detection engine.

6. Future Work:-

Web search engine is a complex System, in which more improvements may be made. There is too much requirement of advanced algorithms to decide what old page be re-crawl and what new one should be crawl. One

more interesting area which require more study is Artificial Intelligence. There is need.

Acknowledgement:-

It gives me immense pleasure in Research on topic Search engine optimization-using crawler optimization. I acknowledge the enormous assistance and excellent co-operation extended by my respected Guide DR.S.K. Gupta, Joint Controller and Examination, M.T.U., Noida.

Reference:-

- [1] Kleinberg , Authoritative Sources in a hyperlinked environment' In Proc 9th ACM SIAM symposium on Discrete Algorithm, 1998.
- [2] R.K. Makhijani, Design of Search engine Using Vector space Model for Personalized search, IJCSMC, Vol2, Issue3, March 2013, Pg 63-66, ISSN 2320-088X.
- [3] Ravi Ranjan, Google search- Anatomy of a Search Engine, VSRD-IJCSIT, Vol2(6), 2012, PG 556-568.
- [4] Dr. S. Saravankumer, A new methodology for search engine optimization without getting Sandboxed, IJARCCCE Vol 1, issue 7, Sep12, ISSN-2278-1021.
- [5] Reinhard O. altro, Proceeding of Internation multiconference of engineers and Computer Scientists, Match 2012, Vol1, HongKong, ISSN-2078-0958, Page 14-16.
- [6] Number of Pages indexed by google overtime, WordPress, 2008.