Web Crawling Methodology for Search Engine Optimization in Web Search Engines

Kalyani Wagaj[1], Subhash V. Pingale[2], M. B. Kulkarni[3]
[1,2]*Department of Computer Science & Engineering, Solapur University, Solapur*
SKN Sinhgad College of Engineering, Korti, Solapur, MS, India
[1]kalyani.wagaj@gmail.com, [2]sub.pingale83@gmail.com

Abstract - Each Search engine's algorithm used for ranking is unique over Internet search engines like Google, Yahoo. For web search engines now a day's SEO strategy is must for online marketing. In this paper, required data for web traffic analysis is used through Google APIs. And methodology includes search engine friendly web based frameworks that commit to improve algorithms over time for page rank calculation, analysis, Web crawling and keyword research. It is kind of SEO tool where basic idea is to increase web traffic for a web site or pages.

Keywords - Web Crawling, Page Rank Calculation, API MOZ and Alexa Ranking.

I.INTRODUCTION

Search engine algorithms take the key elements of a web page, including the page title, content and keyword density, and come up with a ranking for where to place the results on the pages. Each search engine's algorithm is unique, so a top ranking on Yahoo! does not guarantee a prominent ranking on Google, and vice versa.[1] The number of potential results may increase exponentially with the number of sources and links between them. Yet, most of the results may be not necessary especially when they are not relevant to the user. The routing problem, we need to compute results capturing specific elements at the data level. Routing keywords return all the source which may or may not be the relevant sources.[2]

A search engine is a web-based tool that enables users to locate information on the World Wide Web. Popular examples of search engines are Google, Yahoo!, and MSN Search. Search engines utilize automated software applications (referred to as robots, bots, or spiders) that travel along the Web, following links from page to page, site to site. The information gathered by the spiders is used to create a searchable index of the Web.[3]
As a business owner in online marketing there should be proper plan so best ways are available to connect customers.SEO is process of optimising website in such a way that search terms with high popularity leads to a good number of visitors to website. Before planning for SEO strategy to site recent updates of Googles SEO ranking algorithm must be checked out.[4]

1.1. Page Rank Calculation

It is used to find web page popularity in scale point between 0 to 10 to find how reputable a particular web page is according to Google's Ranking algorithm[5] Example for Page Rank Calculation: Assume a small universe of four web pages: **A**, **B**, **C** and **D**. The initial approximation of Page Rank would be evenly divided between these four documents. Hence, each document would begin with an estimated Page Rank of 0.25. In the original form of Page Rank initial values were simply 1. This meant that the sum of all pages was the total number of pages on the web. Later versions of Page Rank would assume a probability distribution between 0 and 1. Here we're going to simply use a probability distribution hence the initial value of 0.25. If pages **B**, **C**, and **D** each only link to **A**, they would each confer 0.25 Page Rank to **A**. All Page Rank i.e. **PR ( )** in this simplistic system would thus gather to **A** because all links would be pointing to **A**.[6]

1.2. Web Crawling

To find out page rank web indexing over Internet is used as web crawlers are web agents which browses world wide web.Google does not leaks the crawling algorithms but web research is carried out to do comparative web data analysis.The process or program used by search engines to download pages from the web for further processing by a search engine that will index the downloaded pages to provide fast searches.[6]
It's important to understand that the search engine user is Google's customer. Once an SEO is able to identify the quality of a website from the searcher's point of view, he's able to build a better brand that attracts both human customers and bots.[7]
Google continues to improve its search engine algorithms through AI and spokespeople for the search engine have continually confirmed that its goal is to "understand" content in relation to user experience, back links, and user behaviour patterns. Instead of ranking for one common factor, the future of SEO should take all factors into consideration. This is, after all, the goal for AI and search engine optimization.In this paper, section II describes the SEO analysis factors.[8]

II. METHODOLOGY

Search engine optimization plays a critical role in web commerce. Without this set of techniques, most websites would be unable to acquire high rankings in searching results. It creates a copy of all the visited pages for later processing by a search engine that will index the

downloaded page to provide fast searches. Checking links or validating HTML code can be used to gather specific type of information from web pages such as harvesting e-mail address (spam). Web crawling is modeled as a multiple queue, single-server polling system on which the web crawler is the server and the web sites are queues. The objective of crawler is to keep the average freshness of pages in its collection as high as possible or to keep the average age of pages as low as possible. To improve the freshness we should penalize the elements that change too often.[5]

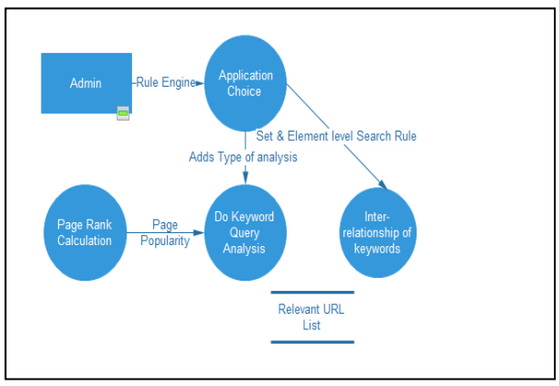

Figure2.1 Work Flow Diagram

A web crawler (also known as a web spider, web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. Other less frequently used names for web crawlers are ants, automatic indexers, bots, and worms. This process is called web crawling. Many sites, in particular search engines, use crawling as a means of providing up-to-date data. 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 website, such as checking links or validating HTML code.[12]

Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).

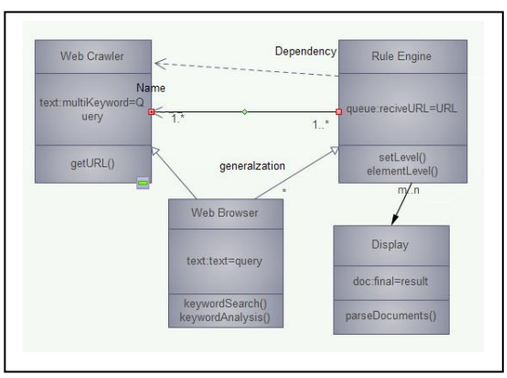

Figure2.2 Class Diagram

A web crawler is one type of BOT, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. Due to dynamic nature of website web pages are up-dated frequently. To keep recent updates or copy of chang-ing web pages, there should be efficient crawling mecha-nism.Crawling process can improve the quality of services provided by search engine.Optimal crawling process and the scheduling algorithm plays a vital role in determining the quality and freshness of web pages. Overall objective is to reduce the search engine embarrassment metrics and to provide best possible search results. It starts with a list of URLs to visit, called the nodes. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of visited URLs, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies[4].

Algorithm for Web Crawler:

Step1.Get the URL
2.Dump the URL into a data structure called queue
3.Go to that URL scan the entire page find out if any links are present, if any URL's are present dump them into that linked list.
4.All the URL's present in the linked list are called as child URL's and the one present in queue are called as parent.
5.Now pick first child URL from linked list dump it into the queue this URL then becomes the parent repeat step 1.
6.Repeat the process for each and every child URL present in the linked list.
7.Keep on doing so till the depth mentioned at the start of the code is reached
Recommendation of all the users will be saved which in turn will be used for endorsement of a specific thing.
This application can be used to create brand awareness on social networking site. This application can even be further integrated with any of the ranking algorithm and the ranking algorithm could be used to rank the data.

III. IMPLEMENTATION

3.1 Prerequisite:

As web crawling is slower as crawler tries to visit every page on website in the crawlers find URL data that is unwanted.To avoid this efficient web crawler should be built with the consideration of following checklist:

1.Crawl depth limit must be set so heavy URL list are not picked.
2.Crawling Exceptions are those part of site Crawler should not visit.
3.Interrupt Pause time before crawler program moving to next page.
4.Save Current Log as crawling may be longer

depending on website depth[6].

3.2 Application Development Environment:

Software Requirements : PHP Code igniter Framework API's by Google,MOZ and Alexa.

Web Hosting and Basic Authentication System as per initial test cases is done.

3.2.1. Module User Interface :Language used PJavascript,HTML5 and CSS3

3..2.2. Core Algorithm Developed: Page Rank and Web Crawling and API'S Key generated for accessing list of URL's.

3.2.3. Report Generation and Download For user: With the help of Code Igniter DOM library

3.2.4. Web hosted and tested for all the services.

Live Work can be accessed at : http://kalyani.net23.net/seopanel/ (More Services are in construction as it is freely hosted loading time will supposed to extend.) Figure shows architecture of the system as deployment (distribution) of software developed.

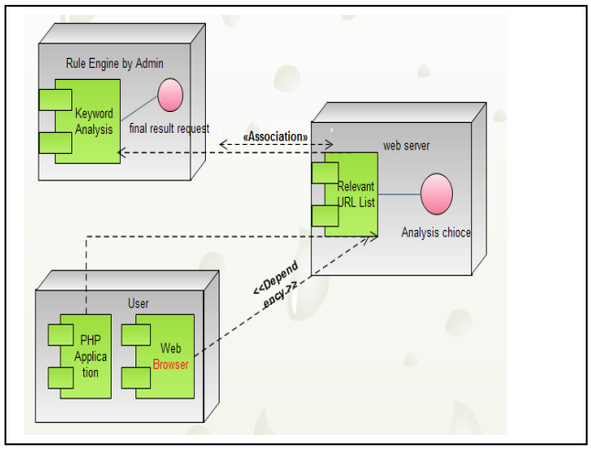

Figure3.1 Deployment Diagram

3.3 Web Crawler and Search Engine Optimization:

When word crawler is used in the context of SEO it is program used by search engines over Internet to analyze the websites.Google's crawler is Googlebot is very active component in SEO marketing Strategy. Building a web crawler is not hard task but choosing right algorithm so that SEO ranking results are improved and traffic is in-creased on website in online business.[4]

Distributed crawling is the process of partitioning of tasks or it is similar to distributed computing technique. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. It basically uses PageRank algorithm for its increased efficiency and quality search. The advantage of distributed web crawler is that it is robust against system crashes and other events. It is more scalable and memory efficient. Also have increased overall download speed and reliability.[5]

Building web crawler to achieve target market in online business is becoming need as many more yet to come on Internet in context of technology. Target market analysis is important as before applying SEO techniques again it requires efficient web crawler to achieve clear and specific analysis results such for link building,SERP analysis and keyword research etc[7].

SEO Link Building	**SEO Tools**	**Crawling**
Done and tested on current web host Platform	Compared with performance for Moz and Alexa Rank	Meta Data Creation and Websites link modification added
Free Web Hosting	Rank Calculation for list of added URLs.	Time for indexing is checked.

Table1.Result Analysis

IV.TEST GOALS

1. The main aim to test this is to insure that:
   - The Proposed system permits only secure and authenticate access.
   - Thus requires the user to enter the URL in correct format.
   - Does all validation time to time as per the need.
   - Takes a single input as user id for the detection of anomalies that is used to generate the recommendations.
   - Does all the ranking calculations internally.
   - Appropriates alerts are generated as per the condition for user convenience.[11]

2. For instance, most searchers don't just simply search once, click on some websites, and be done with it. Instead, they search, click on some websites, edit their search terms, search again, click on some websites, further repeats their search terms, search again, and so on. To avoid such repeating of search procedure the system will provide the optimized search for the query[7]. This system can also be used as efficient method for big data analysis and come up with better solution for SEO techniques used over web analytics.[9]

IV.SCOPE

4.1 Scope as SEO Employee / Job

If you plan to work for some company as a SEO Employee or Consultant then there is not much scope because this field is not just about submitting a website to search engines or web directories but its about understanding how you could really make a website rank into the top and how does all this stuff work. Most people have a idea in their mind that SEO is a very easy job which can be handled easily following which they apply for a placement in different companies to get a quick job.[11]

There is a huge scope in this industry as every web projects needs traffic from the Search, hence everyone is looking for optimizing their websites to rank in the top of the Search Engine Results. If you are looking into the future of this industry then you do not have to worry about its just growing and there is a positive trend for the same. You just have to stick to the work and then you would get success following which you can also shift to developing your own websites and ranking them onto the top instead of working for someone else.[12]

SEO which is known as Search Engine Optimization is the back bone of all the companies of Online Marketing, without which their success in entire world is not possible at any cost.[13]

V. CONCLUSION

In this paper, methodology represented web crawler used can be used in applications as SEO ,search engine optimization is playing vital role in e-commernce, retail and all online business where customers are web site visitors. As SEO analysis accuracy depends on web crawler results so this study helps to build efficient crawler .For job profiles such as SEO anlayst has scope to use page ranking algorithms and web crawling also keeping recent updates for algorithms as parallel with Google's ranking algorithm is becoming challenging task in online marketing In short, now a day's large volume of data on web is generated, to keep the relevant data available as result to user keyword queries, this system will be framework of web analytics and e commerce applications to get the marketplace value on basis of page popularity metric.

ACKNOWLEDGMENT

The author (Kalyani Wagaj) would like to thank Co-author and guide Prof. Subhash Pingale for the valuable support and guidance, Department of Computer Science and Engineering in SKN Sinhgad College of Engineering. This work is supported by College and Solapur University.

REFERENCES

[1] Thanh Tran And Lei Zhang, Keyword Query Routing, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014.

[2] Zhou Cailan,Chen Kai and Li Shasha, Improved PageRank algorithm based on feedback of user clicks, IEEE Wuhan University of Technology, Hubei, China , 2011.

[3] Huang Decai and Qi Huachun, Page Rank Algorithm Research [J], IEEE Computer Engineering, vol. 32, no. 4, pp. 145-146, 2006.

[4] Wang Hui-chang Ruan Shu-hua and Tang Qi-jie, The Implementation of a Web Crawler URL Filter Algorithm Based on Caching, IEEE Computer Science and Engineering, 2009.

[5] Shaojie Qiao, Tianrui Li, Hong Li and Yan Zhu, Jing Peng and Jiangtao Qiu, The Implementation of a Web Crawler URL Filter Algorithm Based on Caching, IEEE Computer Science and Engineering, 2009.

[6] Junghoo Cho and Hector Garcia-Molina, Effective Page Refresh Poli-cies for Web Crawlers, ACM Transactions on Database Systems, 2003.

[7] Satinder Bal and Rajender Nath, A Novel Approach to Filter Non-Modified pages at remote site without downloading during crawling, IEEE International Conference on Advances in Recent Technologies in Communication and Computing 2009.

[8] Joel Coffman, Alfred C. Weaver, "An Empirical Performance Evaluation of Relational Keyword Search Systems", IEEE Transactions on Knowledge and Data Engineering, (Volume: 26 , Issue: 1) Year:2014.

[9] Yen, Shih and Chao "Ranking Metrics and Search Guidance for Learning Object Repository", IEEE Transactions On Learning Technologies, Vol. 3, No. 3, 2010.

[10] http://firstviewonline.com/why-seo-is-important-for-SEO,Ranking and Crawling.

[11] http://www.webseoanalytics.com/free/seo-tools/serp-analysis.php for SEO statics

[12] http://searchengineland.com/guide/what-is-seo for web traffic analysis

[13] http://www.pimall.com/nais/n.engine.html for Search Engine Strategy, Tips And Techniques.