IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SEARCH ENGINE ANALYSIS FOR WEBSITE

[1]Sunil Kumar N, [2]Deepika K

[1]PG Student, [2]Assistant Professor
Master of Computer Applications
R V College of Engineering, Bangalore, Karnataka, India

Abstract: There is a massive size of information present in the system of web pages on the web. So, when any user requests a query, a large volume of search outcomes having different web links matching to the user's query are produced. Out of produced results, only some are appropriate while the remaining are inappropriate. The search engines calculate the significance of a web page by page ranking algorithms. There are smart crawlers available which are automated scripts used for fetching data from the web and allowed to store in databases in pretty formatted manner. Later link is acknowledged and a tree like structures are observed which then defines a page or website strength on the web. This paper discusses determining and implementing different ways to increase the chance of finding a good listing. This will improve the understanding of the usage and its technique in a very practical manner. The search task is performed based on keyword research where popular words and phrases entered by the users in the search engine are used to identify the rank. This gives improved understanding of how high the demand is for certain keywords and how tough it would be to compete for those terms in the search results, contributing some direction to optimization efforts. This gives improved understanding of certain keywords demand and how tough it would be competing through these keywords in the search engine, providing certain ways in optimisation.

Keywords- Page Rank, Keywords, Web Pages, Crawler, In-bound and Out-bound links.

I. INTRODUCTION

The SEO has become tougher and imperative also thanks to the uncountable webpages, hyperlinks if a huge amount of the information concerning the communications of the end-users. This is often the basis why Search Engines plays an important role in today's computer-generated businesses and hence the online page ranking is critical to toughen the simulated presence of any organization [1].

This paper presents an analysis of the page rank method to rank different websites. Search engine analysis considers the algorithm that supports the contents calculation and analysis of the hyperlink structure. Basically, page ranks are obtained using website content, weight and base score of the web page with in-bound links and out-bound links.

Search engine analysis focuses on optimizing factors that are various data which may be structured data, Semi-structured data, and even structured existing on an internet site page. supported it, the analysis also can be broadly classified into the essential three Optimizing Areas [2]:

1.1 Keyword Analysis

Key-word analysis is a key part of a search engine because it would offer you a group of the keywords employed by the end-users from the domain. With this user can change relevant keywords on their web site pages to boost the virtual presence of the website. A keyword analysis course during which we learn the keyword for our website by means of the ad word. Using this ad word, we realize the opposition and search section for our keyword. we will get a better understanding of keywords certainly through the search result. The keywords are categorized into three types shot tell, long tell, and related [3].

1.2 On-Page Optimization

This method includes the introduction of the keywords in keywords in alt tag, keyword density, title tag, keyword within the meta tag etc. This is recurrently a real benefit followed by every sensible webmaster so as to form the online page more efficiently in terms of appearance and rank. By applying this system, page rank definitely improves within the program and simultaneously gives far better fulfilment to the visitor because the results are what the end-user was expecting.

On-page optimization is very frequently monitored using different ways to form a webpage; for instance, it is done through changing or modifying the title, body text, URL, and density of the keywords throughout a webpage. Though, one must look out for the newest updates by search engines in respect of the keyword's density. If the on-page optimization is carried out with care, it is sure to produce unbelievable results and productive traffic on the website page. On-page optimization is skilful enough to be more explicit in terms of the visitors going to see the website [4].

1.3 Off-Page Optimization

Off-page optimization is currently becoming more and stronger because it must support with the present Google updates and therefore the clever use of the backlinks. This optimization focuses more on forwarding a user from other webpages with the support of a backlink. The websites structured by more backlinks are likely to urge successful and resultant visitors. However, care has got to be taken while Link-building for websites

We find various thoughts declaring that backlinks are almost outdated today but it's not the backlinks, but it is the Back-link plan fails because mainly due to the placing a link without giving a clue to the entire plan and link structure. What makes the difference is that the application of the backlink with the redirection of page through text on page or a blog describing the importance of renewable sources of energy to a web page. Off-page optimization are done through back link submission, Social media management, blog submission, website sharing, Social media bookmarking site, etc.

II. RELATED WORK

This paper [5] shows the effectiveness of search engine and keyword-based search that's incorporated in traditional information retrieval systems. Also explains a technique which can be used for effective information retrieval and real user experience for Search Engine. Paper includes the Search Engine's construction from request engine to ranking of Web Pages. The ability of search engines with indexing of site pages over query optimizer technique by author in this paper.

This paper [6] defines the function of selected Search Engine Optimization method to a recently created website throughout the entire website lifespan, from start, through growth, and finally to commence site optimization. Visitors, Pageviews and Ranking are components for efficiency of the SEO techniques. It was found that there was a drastic increase in the rankings of search engines and the mass of users who visited the site also increased. It showed a positive effect by increasing the traffic on the Google search rankings.

This paper [7] focuses on the platform called World Wide Web where the structure of web pages is accessible in heaps of web data. So that whenever a user defines a query, huge search results bearing a variety of web links equivalent to the user's query are developed. As Google innovated Page Ranking Algorithm which has been benchmarked as a standard algorithm since there exists no algorithm for the similar process of ranking web pages. Immediately following a figure of factors like numbers or weights of visit of a web link is included in standard page rank algorithm.

This Paper [8] shows the process of search engine valuation which is based on page level keywords. The keywords originated on distinct pages of websites are called as page level keywords. Page level keywords include different tags and URL string in the HTML. Our system repossessed over keywords from the site pages and compared with most searched keywords by search engines.

This paper [9] proposes a novel technique which makes user search data quite efficient. This method gives a connection between the searched document and user request. It is also considering the web link structure and user query content. Using deep web analysis of site pages and relevance of documents our system can provide fast query results.

This paper [10] proposes an improved technique to rank an internet page from a group of given sites. System is using semantic similarity algorithms alongside AI techniques to rank the webpages. System is tested on thousand sites and comes under various categories like education, computer, programming, chemistry etc. Systems are often further improved by implementing on the cloud servers and by using multithreading techniques to enhance the time efficiency. Systems are often further checked by increasing the amount of website categories. Multithreading techniques also can be integrated within the proposed system to enhance the general performance of the system

III. PROPOSED METHODOLOGY

The implemented methods were not able to wrap up all the functionalities. The proposed system uses web structure, data and program optimization techniques to analyse the importance of each page in a large dataset of web structure. Page Rank is obtained by allotting a mathematical weight to a web page as per its virtual importance. The analysis is done through a keyword analysis process which helps in bringing the visitors to your website. The detection of links that either send the visitors to a web page that no longer exists or that do not work or isn't available for the given moment is all analysed by a broken link analyser. Hence the proposed system is a unique combination of all these modules, while mainly focusing on increasing the page rank for a given website.

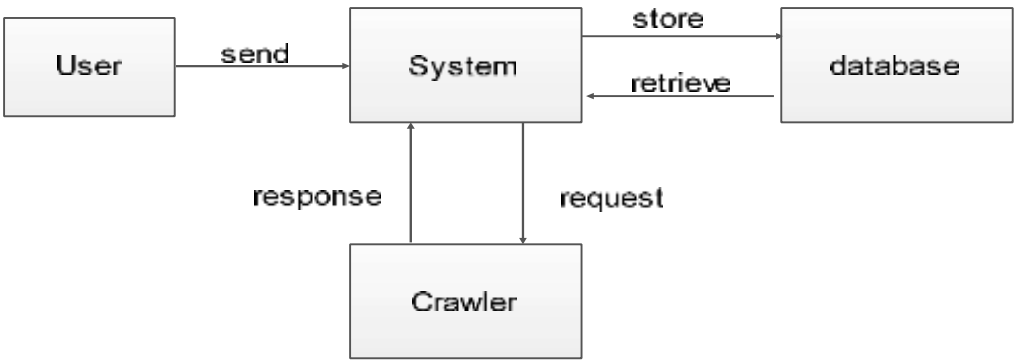


fig 1: working of the proposed system

As shown in figure 1 the enters a URL of the desired website, this is the input to the developed application, further the application requests the crawler to gather data from the web. This data is stored into the database of the application which is used by the application to perform the functionalities like allocating the page rank, finding the broken links by using hyperlinks and keyword analysis.

IV. RESULT

fig 2: dashboard of the application

Figure 2 shows the dashboard of the application with the input to crawler is given in two Types that is website URL and number of pages to be crawled. In this application the BeautifulSoup web crawler is used to parse html and xml documents which is useful for extracting data from web pages

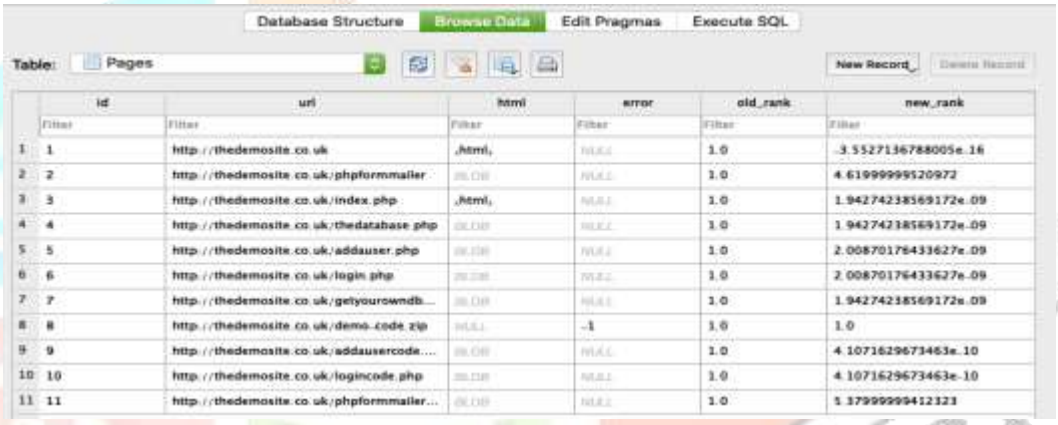


fig 3: retrieved data from crawler

Figure 3 shows the data retrieved by the crawler and collection to the local system with the help of a lightweight SQLite database. The table contains primary id for each hyperlink and complete html content with all tags. further the link structure and allocation of initial rank for all the pages to value one. The system provides a unique ID for each inbound link and outbound links and the web structure of the website is analysed and structured by beautiful soup. This Structure is further used in distribution of ranks in page rank algorithm by calculating out-bound links.

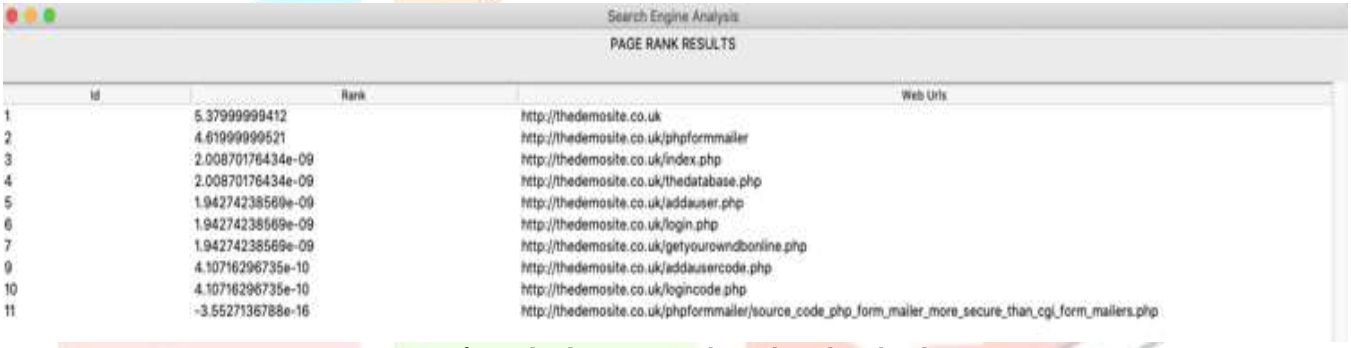
fig 4: displays the keywords on site with count

Figure 4 shows the total number of key words by using natural language processing to find out word count and also most repeated phrases in title tag, paragraph tag and heading tag. Each page is analysed, and results are produced which helps in modifying the ranked words by search engines.

fig 5: displays list of all broken links of website

Figure 5 shows broken link analysis where websites don't need dead links because it provides poor ranking of the website. This application is designed to identify the broken links like

- 400 - Bad Request
- 401- Unauthorized
- 403 - Forbidden
- 404 - Not Found
- 500 - Internal Server Error
- 502 - Bad Gateway
- 504 - Gateway Timeout



Id	Rank	Web Urls
1	5.37999999412	http://thedemosite.co.uk
2	4.61999999521	http://thedemosite.co.uk/phpformmailer
3	2.00870176434e-09	http://thedemosite.co.uk/index.php
4	2.00870176434e-09	http://thedemosite.co.uk/thedatabase.php
5	1.94274238569e-09	http://thedemosite.co.uk/addauser.php
6	1.94274238569e-09	http://thedemosite.co.uk/login.php
7	1.94274238569e-09	http://thedemosite.co.uk/getyourowndbonline.php
9	4.10716296735e-10	http://thedemosite.co.uk/addausercode.php
10	4.10716296735e-10	http://thedemosite.co.uk/logincode.php
11	-3.5527136788e-16	http://thedemosite.co.uk/phpformmailer/source_code_php_form_mailer_more_secure_than_cgi_form_mailers.php

fig 6: displays page rank result with web url

Figure 6 shows implementation of Iterative method to calculate the page rank corresponding to web link structure. The result of the iterative method shows the rank value and web URL with high weightage and a greater number of times this method is executed it provides closer to final value.

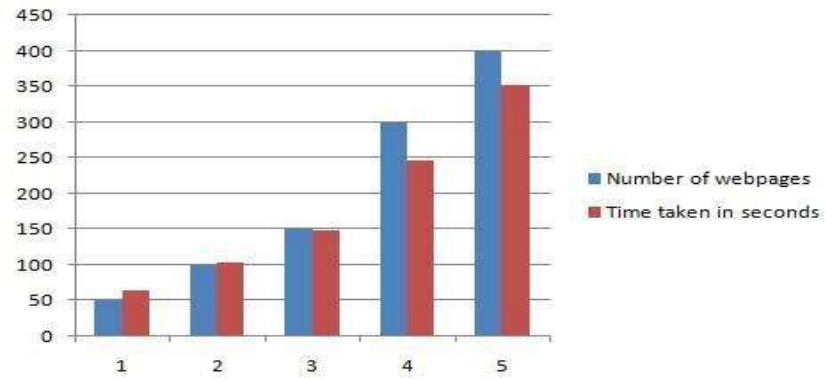


fig 7: status of time taken by crawler to crawl number of pages

Figure 7 represents the graph for the time taken to crawl 50, 100, 150, 300 and 400 webpages.

V. CONCLUSION

Due to the availability of a huge number of webpages, the search engine has a very important job of providing the related pages to the user, most Search Engines use Page Ranking Algorithm to rank web pages according to the excellence of page content and their presence over the internet. The proposed method is a platform for increasing the page rank by acquiring the following properties like ranking ability, index speed, important links, elimination from bad links. This system also provides ranking for every web page by applying Page Ranking Algorithm. Page Rank is obtained by allotting a mathematical weight to a web page as per its virtual importance. The analysis is done through a keyword analysis process which helps in bringing the visitors to your website. The detection of links that either send the visitors to a web page that no longer exists or that do not work or isn't available for the given moment is all analysed by a broken link analyser. Hence the proposed system is a unique combination of all these modules, while mainly focusing on increasing the page rank for a given website.

VI. REFERENCES

[1]. Dr.Divya Gupta, Devika Singh, "User Preference Based Page Ranking Algorithm", ICCCA 2016, ISBN: 978-1-5090-1666-2, IEEE, 2016.

[2]. Shalini Sharma," Web Crawler", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014 ISSN:2277 128X.

[3]. S.Amudha, "Web Crawler For Mining Web Data", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 02 | Feb -2017 p-ISSN: 2395-0072.

[4]. Mini Singh Ahuja, Dr Jatinder Singh, Bal Varnica, "Web Crawler: Extracting the Web Data", International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014.

[5]. Pratiksha sahiwal, S.R.Tandan, Priyanka Tripathi, "Web Information Retrieval Using Python and BeautifulSoup", Volume 4 Issue VI, June 2016, ISSN: 2321-9653.

[6]. Zhou Hao, Pu Qiumei , " An Improved PageRank Algorithm Based on Web Content", ISBN: 978-1-4673-6593-2, IEEE, Aug 2016.

[7]. Junghoo Cho, Hector Garcia-Molina, and Lawrence Page," Efficient Client crawling through url ordering". Proceedings of the twelfth International Web Conference, May 2010.

[8]. Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.

[9]. Chongchong Zhao, Zhiqiang Zhang, Hualong Li, Xiaoqin Xin, "A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering", ISBN: 978-1-42448728-8/11/- 2011 IEEE.

[10]. B.S Pooja Sharma, St Deepak Tyagi, "Weighted Page Content Rank for Ordering Web Search Result", International Journal of Engineering Science and Technology, vol. 2, no. 12, pp. 7301-7310, 2010.