A New Ranking Technique for Ranking Phase of Search Engine: Size based Ranking Algorithm (SBRA)

Prabhat Kumar Singh
Asst. Professor
TMU, Moradabad

Gaurav Agarwal
Asst. Professor
KIMT, Moradabad

Sachi Gupta
Asst. Professor
KIMT, Moradabad

ABSTRACT

The objective of this paper is to propose a new ranking technique or method for the web page rank. Different search engines use the different algorithm for the ranking of the web pages. Web Page ranking algorithm works only on the web page repository or indexed pages to the search engine for ranking. In reality, search engines works in two phases one is crawling phase and another is ranking phase. In this paper proposed work is based on the ranking phase.In this Paper designed new ranking algorithm which will be known as Size Based Ranking (SBR) Algorithm. This algorithm is cover more area of the web pages and hope that this algorithm is provide the best, accurate and sufficient data or information or matter to the user according to the need.

KEYWORDS

Density of keyword on upper half page, Density of keyword on lower half page, Number of successors to the page, Freshness value of the page.

1. INTRODUCTION

Mammoth data are available on the web or internet which is used by the user. World Wide Web contains the web pages on the internet. WWW provide us with great amounts of useful information electronically available as hypertext. A vast number of web pages are continually being added every day, and information is constantly changing [1].

Search engines are computer programs that travel the Web, gather the text of Web pages and make it possible to search for them. Please keep in mind that no search engine covers the entire Web. In fact, experts estimate that the largest search engines cover only 15% of the World Wide Web [2].

A Search Engine is a web site that collects and organizes content from all over the internet. Those wishing to locate something would enter a query about what they'd like to find and the engine provides links to content that matches what they want. Creating a Search Engine which scales even to today's web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second [3].

2. WORKING OF SEARCH ENGINE

The researchers study some papers about the search engines and found the Search Engine work on four modules which names are following and all module is shown in figure1with discription:-

1. Crawler module
2. Indexer module
3. Query module
4. Ranking module

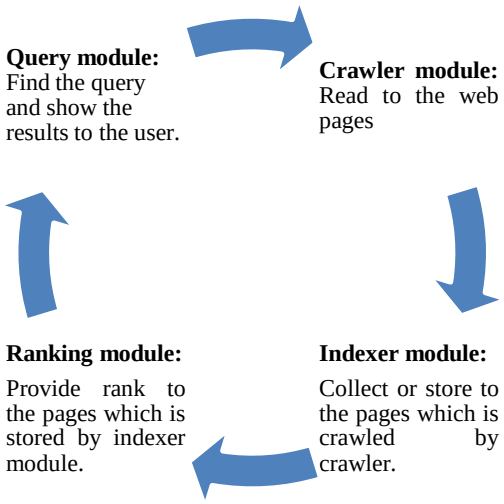

Figure1: Shows the Search Engines Module in Cyclic Process

2.1 Crawler Module

Search the URL on the web and linkages page with on single URL. Crawler behavior is depending on following policies [1]:-

i. **Selection Policies:** - In this policy we decided that which pages is downloaded or which is discarded.
ii. **Revisited Policies: -** In this policy we decided that which page is revisited for the changes.
iii. **Politeness Policies**: - In this policy that states how to avoid overloading
iv. **Parallelization Policies:**-That states how to coordinate show web crawlers

2.2 Indexer Module

Extracts the words from each page it visits and records URL's. Its results into a large lookup table that gives a list of URL pointing to pages where each word occurs. The table list of those pages which were covered in crawling process.

2.3 Query module

Query module of a search engine receives search requests from users in the form of keywords or content or phrases.

And search the keywords or content or phrases in the own data base.

2.4 Ranking module

Many algorithms are available for ranking to the web pages. In this module we sort the results according to ranking algorithm. Different ranking algorithm is defined the ranked of pages. Some approaches are following:-

1. Top –down approach or parsing
2. Bottom –up approach or parsing

3. RELATED WORK

The purpose of Page Ranking is to measure the relative importance of the pages in the web. There are many algorithms for this purpose [4]. The most important page ranking algorithms are describes below:

- Hyper Search,
- Hyperlink-Induced Topic Search (HITS),
- PageRank,
- TrustRank, and
- OPIC

3.1 Hyper Search Method

Hyper Search has been the first published technique to measure the importance of the pages in the web. This algorithm served as a base for the next ones. For more information about Hyper Search refer to [5].

3.2 Hyperlink-Induced Topic Search Method

HITS algorithm, also known as Hubs and Authorities, is a link analysis algorithm for the web. It is executed at query time and is used to modify the ranking of the results of a search by analyzing the link structure of the pages that will appear in the result of the search. HITS algorithm assigns two different values to each web page: its authority value, and its hub value. The authority value of a page represents the value of the content in the page; meanwhile the hub value estimates the value of its links to other pages.

The first step in the HITS algorithm is to retrieve the set of pages in the result of the search, as the HITS algorithm only analyzes the structure of the pages in the output of the search, instead of all the web pages [8].

3.3 PageRank Method

PageRank is a link analysis algorithm to measure the page relevance in a hyperlinked set of documents, such as the World Wide Web. This algorithm assigns a numerical weight to each document. This numerical weight is also called Page Rank of the document. The Page Rank of a web page represents the likelihood that a person randomly clicking will arrive at this page. The Page Rank algorithm requires several iterations to be executed.

At each iteration, the values will be better approximated to the real value.In its simplest form, Page Rank uses the next formula for each web page at each iteration:

$$PR(u) = \sum_{v \in Bu} \frac{PR(v)}{L(v)}$$

Where u is a web page, Bu is the set of pages that link to u, PR(v) is the PageRank of v, and L(v) is the number of out-links in page v [7].

3.4 TrustRank

TrustRank is a semi-automatic link analysis technique for separating useful WebPages from spam. The HITS and the PageRank algorithms can also be used for this purpose, but They have been subject to manipulation. To avoid this manipulation, the TrustRank algorithm selects a small set of documents. This set of documents will be evaluated by an expert. Once the reputable seed pages are manually identified, a crawl extending outward from the seed set seeks out similarly reliable and trustworthy pages. For more information about TrustRank, please refer to [6].

3.5 On-line Page Importance Computation (OPIC)

The On-line Page Importance Computation algorithm, also known as OPIC, is a link analysis algorithm. Its capability to compute the page importance without storing the whole links graph is what makes it different from other algorithms.

The OPIC algorithm keeps 2 values for each page: its cash, and its history. Initially, some cash is distributed uniformly among all the nodes, for the case of N nodes, 1/N cash will be assigned to each node. The cash of a web page contains the sum of the cash obtained by the page since the last time it was crawled. The history of the page contains the sum of the cash that the page has obtained since the algorithm started until the latest time it was crawled. When a page p is retrieved, its cash is added to its history [9].

4. PROPOSED WORK

In this paper the thesis work and designed algorithm will be provide the fastest way to find the rank of web pages which will be known as Size Based ranking Algorithm (SBRA). Because this algorithm is used two different approaches which are applied on a single page at the midpoint of web page which due to search of content is fastest and it will be take just half time compared to previous required time to searching the content of web pages. In SBRA, used the four variables to design this algorithm which are following and also describe by graph which is shown in figure2:-

1. Density of keyword on upper half page,
2. Density of keyword on lower half page,
3. Number of successors to the page,
4. Freshness value of the page.

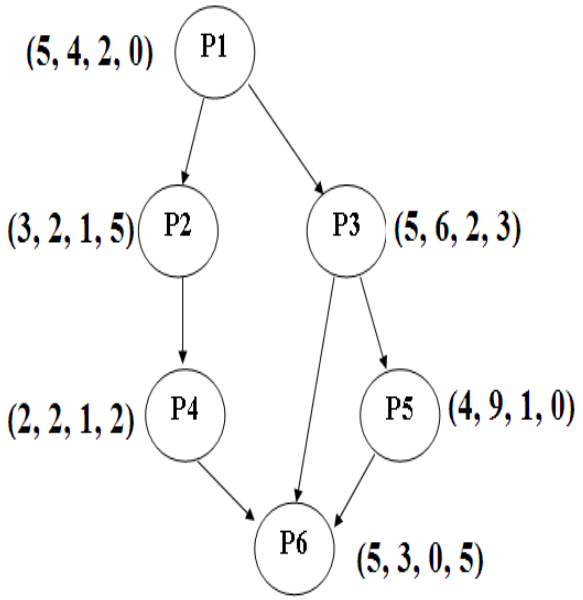

Figure 2: Shows Page Graph with Density of keyword on upper half page, Density of keyword on lower half page, Number of successor to the page and Freshness value of the web page.

5. SIZE BASED RANKING ALGORITHM (SBRA)

SBR Algorithms works on the two different approaches one is Top-Down approach and another one is Bottom- Up approach for calculating rank of the web page. A SBR Algorithm is based on the size of the web page and its working start from the middle of the web page. A web Page is divided in to two equals' half part. Then Bottom - Up approach applying in Upper Half Page and Top-Down approach applying in Lower Half Page. This algorithm is used Upper Half Density of Keywords and Lower Half Density of Keywords and Number of outgoing links from a single web page. Calculate two different ranks one for Upper Half Page and another for Lower Half Page then both ranks are add then find the total rank of web page. Results indicate that the proposed method is very fast, feasible, more accurate and valuable in ranking of the web pages. SBR Algorithm working process is shown in figure 3.

SBR Algorithm is based on the four variable which is 'density of keywords in upper half page', 'density of keywords in lower half page', 'successor of the page', 'freshness value of the page'. If multiprocessor system is available then compute Rank on different processors for Upper Half (UH) and Lower Half (LH). Web Page Rank formula is following which is used in this ranking algorithm.

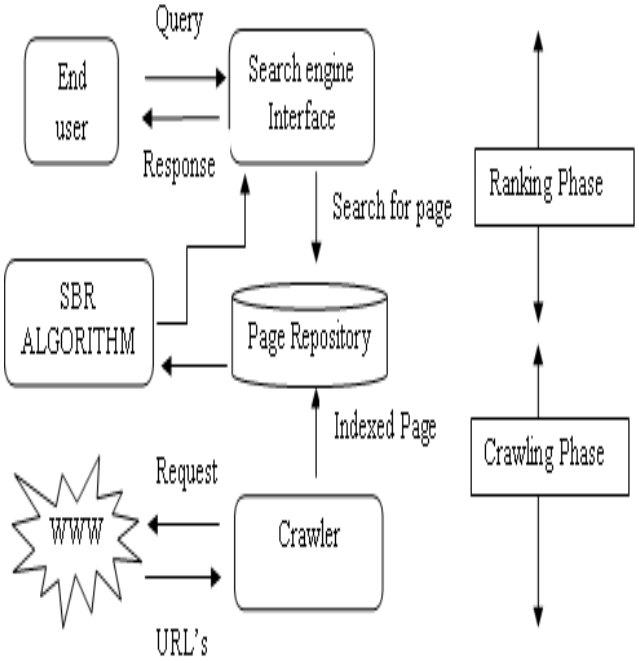

Figure 3: SBR Algorithm placed in the search engine working model

a) For processor 1

Rank P_e = *(Density of keywords in upper half + Freshness value) / 100;*

Otherwise

Rank P_{other} = *(Density of keywords in Upper half + Freshness Value+ Maximum rank Value of successor) / 100;*

b) For processor 2

Rank P_t = *(Density of keywords in lower half + Freshness value) / 100;*

Otherwise

Rank P_{other} = *(Density of keywords in lower half + Freshness value +Maximum* rank *Value of predecessor) /100;*

6. EXPERIMENT AND RESULT ANALYSIS

6.1 Comparison of SBR Algorithm with Google Page Rank Algorithm:

The researchers have an experimental setup for result analysis of SBR Algorithm and done online comparison with the Google Page Rank Algorithm using the website markhorrel.com. The researchers got some value for web pages by Google Page Rank Algorithm and SBR Algorithm which is shown in Table 1. And produce the comparison Graph and Pi chart for PG1 from Table 1. Thus also produce the graph for PG2, PG3, etc.

Table 1: Value of diffrent Pages using Google Page Rank Algorithm and SBR Algorithm

	PG1		PG2		PG3		PG4		PG5	
	GPR	**SBR**	**GPR**	**SBR**	**GPR**	**SBR**	**GPR**	**SBR**	**GPR**	**SBR**
P1	0.150	0.333	0.150	0.323	0.150	0.393	0.150	0.263	0.150	0.413
P2	0.176	0.454	0.171	0.432	0.171	0.345	0.176	0.403	0.182	0.364
P3	0.180	0.354	0.150	0.422	0.174	0.454	0.150	0.373	0.336	0.556
P4	0.201	0.362	0.171	0.363	0.196	0.285	0.219	0.495	0.245	0.485
P5	0.180	0.623	0.342	0.305	0.221	0.385	0.225	0.415	0.192	0.255
P6	0.205	0.433	0.309	0.264	0.19 6	0.365	0.308	0.465	0.359	0.605
P7	0.226	0.365	0.330	0.415	0.271	0.374	0.344	0.275	0.192	0.383
P8	0.323	0.326	0.269	0.485	0.380	0.375	0.569	0.432	0.439	0.463
P9	0.715	0.316	0.686	0.353	0.903	0.604	0.332	0.455	0.507	0.355
P10	0.976	0.572	0.530	0.543	1.101	0.384	0.646	0.412	0.736	0.502
(+)	**3.331**	**4.137**	**3.109**	**3.906**	**3.762**	**3.963**	**3.117**	**3.987**	**3.338**	**4.382**

Comparison Graph is showing the analysis of results based on the comparison of PG1from value of Table 1and this comparsion graph is shown in figure 4.

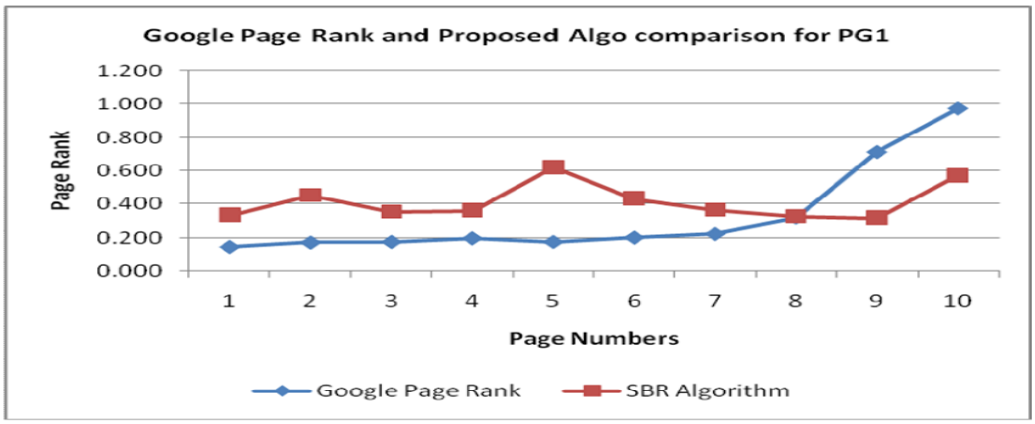

Figure 4: Comparison of Google Page Rank Algorithm and SBR Algorithm for PG1

Pie chart is based on the comparison graph which is showing the covered area by Google Page Rank and SBR Algorithm as shown in figure 5.

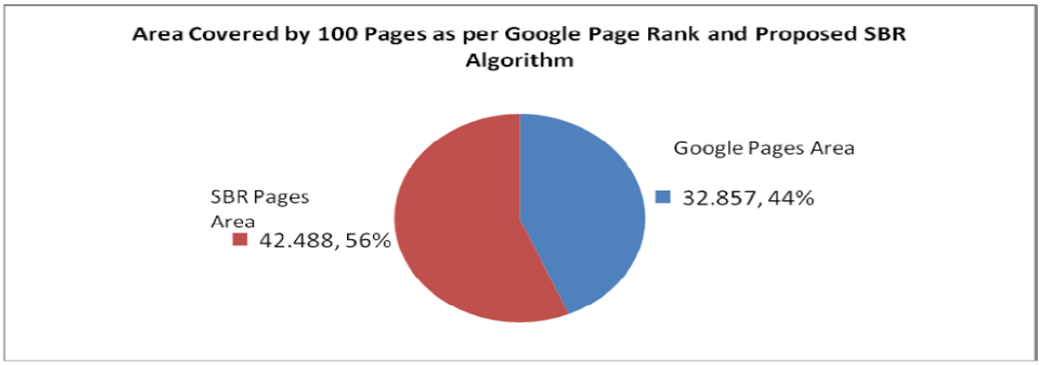

Figure 5: Area Covered by Google Page Rank Algorithm and SBR Algorithm

7. CONCLUSION

Standard web searchers show very little information than what the user want. The page ranking should give meaningful information to the user, thus making a more usable to search engine.

Previous ranking algorithms are based on links or hub data or different things but not the content. Content is king but link is not. While link is play very important role to finding page rank. These ranking algorithms are suitable for the finding to web page ranking but these are not so fast to find the best results compare to SBR Algorithm.

In SBR Algorithm for Search Engine Page Ranking, enrooted an algorithm, in which used four range variable its means the upper half density of keywords, the lower half density of keywords, number of successors to the web pages and the freshness value of the web page. Density shows the occurrence of the keyword on the particular web page. Numbers of successors represent the outgoing link to a single web page. Freshness value is the age of the web page.

In SBR algorithm, two density of keywords are used one for upper half and second for lower half which due to performance of page rank is very high and Many large Web crawlers start from initial pages, fetch every links from them, and continually repeat this process without change any policies that help them to better crawling and improving performance of those. There are obviously many improvements to the Algorithm that can be made.

In future work, the same algorithm can be implemented for using genetic algorithm so that on the basis of parameter like mutation crossover the efficiency and the effectiveness of the algorithm may be improved. As for this proposed work , SBRA used top to down and bottom to up approach for the calculation of the page rank there can be many alternative policies for the same.

Finally hope that the SBR Algorithm is working very fast compare to another Algorithm. SBR Algorithm is covered more area compare to Google Page Rank algorithm. It's providing the best and fast rank to the web pages because two different approached is worked at a time. And it is used for future prospectus in generic algorithm.

8. ACKNOWLEDGEMENT

I would like to express my deep gratitude and respect to Mr. Gaurav Agarwal whose advices and insight was valuable to me. For all I learned from him, and for providing the Vision Lab for the experiments. I would also like to thank him for being an open person to ideas, and for encouraging and helping me to shape my interest and ideas and for his continuous help and support in all stages of this thesis.

9. REFERENCES

[1] S.S. Dhenakaran1 and K. Thirugnana Sambanthan2 "web crawler - an overview "International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267.

[2] Linda Fortney Montgomery College Rockville Campus Library "Web search engines" la 634 aug. 2003.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hyper textual web searchengine. In Computer Networks and ISDN Systems, pages 107-117, 1998.

[4] Pau Valles Fradera" Personalizing web search and crawling from clickstream data"19/01/2009.

[5] Massimo Marchiori. The quest for correct information on the web: Hyper search engines. In Proceedings of the Sixth International World Wide Web Conference (WWW6), 1997.

[6] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam withtrustrank. Technical Report 2004-17, Stanford Info Lab, March 2004.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hyper textual web search engine. In Computer Networks and ISDN Systems, pages 107-117, 1998.

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. In 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[9] Serge Abiteboul, Mihai Preda, and Gregory Cobena. Adaptive on-line page importance computation. In WWW '03: Proceedings of the 12th international conference on WorldWide Web, pages 280-290, New York, NY, USA, 2003. ACM.