A New Ranking Algorithm for Search Engine: Content's Weight based Page Ranking

Charanjit Singh
Research Scholar
Guru Kashi University
Talwandi Sabo

Vijay Laxmi, PhD
Professor and Dean
Guru Kashi University
Talwandi Sabo

Arvinder Singh Kang, PhD
Professor and Dean
Chandigarh University
Gharaun

ABSTRACT
The objective of this paper is to propose a new ranking technique or algorithm for web page rank. Different search engines use various ranking algorithms 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 work in two phases one is crawling phase and another is ranking phase. This proposed Algorithm is based on the ranking phase. In this Paper purposed a new ranking algorithm which will be known as Content's Weight Based Page Ranking Algorithm (CWPRA). This algorithm Calculate the Page Rank based upon the weight of contents which is related to user query. Hope this algorithm provide the best, accurate and sufficient data or information or matter to the user according to the need.

Keywords
Web Mining; Search Engine; Page Ranking

1. INTRODUCTION
We Massive data are available on the web or internet which is used by the user. World Wide Web contains the web pages on the internet and provide us with great amounts of useful information electronically available as hypertext. A huge number of web pages are continually being added every day, and information is regularly 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. No search engine covers the entire Web, moreover experts estimate that the largest search engines cover only 15% of the World Wide Web [2].

2. WORKING OF SEARCH ENGINE
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. [3]

The researchers study some papers about the search engines and conclude that the Search Engine work in mainly four Phases which names are following and all

phases are connected or transfer, manage the data are shown in figure1 with description: -

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

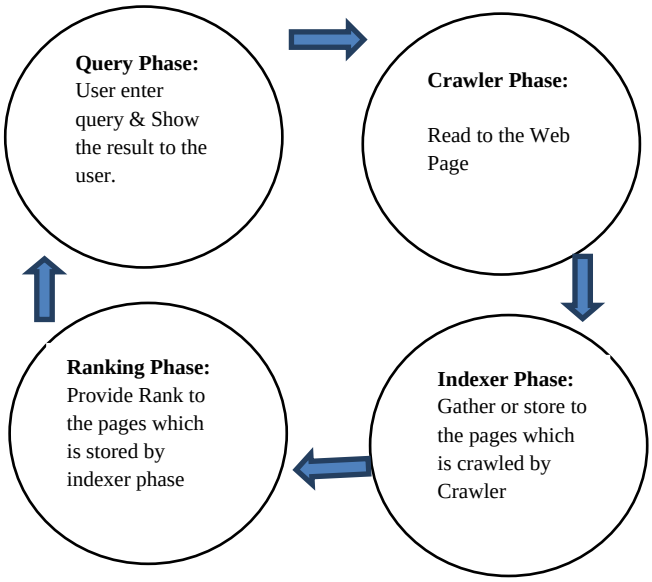

Figure1: Shows the Search Engines Phases in Cyclic Process

2.1 Crawler Phase
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. [3]

Search the URL on the web and linkages page with on single URL. Crawler behavior is liable on following policies [4]: -

(a) Selection Policies: - In this policy we decided that which pages is downloaded or which is discarded.

(b) Revisited Policies: - In this policy we decided that which page is revisited for the changes.

(c) Politeness Policies: - In this policy that states how to avoid overloading

(d) Parallelization Policies: - That states how to coordinate show web crawlers

2.2 Indexer Phase
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 Phase
In Query Phase of a search engine, it receives search query from users in the form of keywords or content or phrases and then search in the own data base.

2.4 Ranking Phase

Numerous algorithms are offered for ranking to the web pages. In this module, the results sorts 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 Algorithms is to measure the relative importance of the pages in the web. There are many algorithms for this purpose [5].

The most important page ranking algorithms are describing below:

- Hyper Search,
- Hyperlink-Induced Topic Search (HITS),
- PageRank,
- Weighted Page Rank 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 [6].

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 [7].

3.3 PageRank Method

Page Rank is used to measure the importance of website pages by counting the number and quality of links to a page. This algorithm states that the Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it (incoming links). If a page has some important incoming links to it than its outgoing links to other pages also become important. A page that is linked to by many pages with high Page Rank receives a high rank itself. [8]

The Page Rank algorithm requires several iterations to be executed. [4] 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 outlinks in page v [7].

3.4 Weighted Page Rank Algorithm

An extension of PageRank algorithm and assigns rank values to pages according to their importance or popularity rather than dividing it evenly. [8] The popularity is assigned in terms of weight values to incoming and outgoing links and are denoted as Win(v, u) and Wout (v, u) respectively. Win(v, u) is the weight of link (v,u) calculated on the basis of incoming links to page u and the number of incoming links to all reference (outgoing linked) pages of page v.

$$W^{in}_{(v,u)} = I_u \,/\, \sum_{P \in R(v)} I_P$$

where Iu and Ip represent the number of incoming links of page u and page p, R(v) is the reference page list of page v. Wout(v,u) is the weight of link (v,u) calculated on the basis of the number of outgoing links of page u and the number of outgoing links of all the reference pages of page v.

$$W^{out}_{(v,u)} = O_u \,/\, \sum_{P \in R(v)} O_P$$

Here Ou and Op represents, the number of outgoing links of page u and page p, respectively. Then the weighted Page Rank is given by a formula:

$$WPR(u) = (1-d) + d \sum_{V \in B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

3.5 On-line Page Importance Computation (OPIC)

A link analysis algorithm OPIC (On-line Page Importance Computation) is used to compute the page importance without storing the whole links graph is what makes it different from other algorithms. 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 proposed work the new algorithm is designed that will be provide the faster way to find the rank of web pages which will be known as Contents Weight based Page Ranking Algorithm (CWPRA). Because this algorithm is used two different approaches which are applied on the Huge database of a URLs (Web Pages) which is on another exclusive web Address (Remote Server). Two different ranking algorithms are used weighted content Ranking which find & rank the web page having content related to user's query on web page and Page Ranking algorithm calculates and assign rank to web pages on the basis of in-link and out-links to respective web page.

4.1 CWPR Algorithm

Content's Weight Based Page Ranking Algorithm

1. User enter key word for search
2. X= {link1, link2, lik3, linkn}
3. Initialize k=0 and n= X.length
4. For each k<n
5. If content is find in kth index the flag=1;
6. Initialize i=0 i<kth index document, counter=0;
7. If keyword is find, then counter++;
8. If i>k, return to step 4
9. If k>N, return false
10. The set the rank of link using below formula

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

a. PR(A) is the PageRank of page A where A is counting of content

b. PR(Ti) is the PageRank of pages Ti which link to page A,

c. C(Ti) is the number of outbound links on page Ti

d. d is a damping factor which can be set between 0 & 1.

In Proposed Algorithm CWPRA, when a user searches a keyword 'x' in huge database of urls. Firstly, it will check the frequency of keyword covered in the various urls which is fetched from particular domain and update the flag value on every hit. Afterword, it will count the number of keywords founds in the urls and stored in array according to its frequency occurred in urls.

From this on words, second algorithm works which define the ranks of urls on the basis of in-links and out-links with pre-defined formula.

So according to CWPRA, we can assign the higher rank to the web pages which has highest frequency of user's query keyword. Higher rank's url will be on the top of result list according to their ranks.

5. CONCLUSION

Normal web searchers show limited information than what the user want in the query. Various hybrid page ranking algorithms 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 various things but not the content. Content is more important than the link. While link analysis is play very important role to finding page rank. These ranking algorithms are suitable for the finding to web page ranking but may not so fast to find the best results compare to CWPR Algorithm.

In Content's weight based page rank algorithm may use to find the useful links according to the user's query. This algorithm will work on two principles, firstly it will check the content of web page and on the behalf of the fetched list of URLs from the database it will compute their page ranks.

In future work, the same algorithm can be implemented for researcher's research, so that on the basis of parameter like web contents weight and page rank of web pages, the efficiency and the effectiveness of the algorithm may be improved.

Finally hope that the CWPR Algorithm is working very fast compare to existed Algorithms. CWPR Algorithm is covered more area compare to Google Page Rank algorithm. It'll provides 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 web search engines.

6. ACKNOWLEDGEMENT

I would like to express my deep thankfulness and respect to Dr. Vijay laxmi whose advices and insight was valuable to me. For all I learned from her. I would also like to thank her 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 research work.

7. REFERENCES

[1] S.S. Dhenakaran and K. Thirugnana Sambanthan "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] Singh, C., Laxmi, V. and Singh, A. "Evolution of web Mining" National Conference of Science and Technology (NCAST-2016) at RIMT-IET, MandiGobindgarh, March 2016.

[4]. Singh, Parbhat Kumar., Agarwal, G. and Gupta, S. "A New Ranking Technique for Ranking Phase of Search Engine: Size based Ranking Algorithm (SBRA)", International Journal of Computer Applications (0975 – 8887), Vol-85-No -5, Nov-2013.

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

[6]. 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.

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

[8] Singh, C. and Kautish, S. "Page Ranking Algorithms for Web Mining: A Review" International Journal of Computer Applications (0975 – 8887), International Conference on Advancements in Engineering and Technology (ICAET 2015).

[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.