*International Journal of Information Technology and Management*

*Vol. VI, Issue No. I, February-2014, ISSN 2249-4510*

# A COMPARATIVE ANALYSIS ON THE IMPROVEMENT OF GOOGLE'S WEB PAGE RANKING ALGORITHMS

International Journal of Information Technology and Management
Vol. VI, Issue No. I, February-2014, ISSN 2249-4510

# A Comparative Analysis on the Improvement of Google's Web Page Ranking Algorithms

**Ms. Rubina Khan**

Asst. Prof. Immersive Institute of Technology

***Abstract – This paper presents different parallel implementations of Google's PageRank algorithm. The purpose is to compare different methods for computing PageRank on large domains of the Web. The iterative algorithms used are the Power method and the Arnoldi method.***

***We have implemented these algorithms in a parallel environment and created a basic Webcrawler to gather test data. Tests have then been carried out with the different algorithms using various test data. In this article, we introduce the Google's method for quality ranking of web page in a formal mathematical format, use the power iteration to improve the PageRank, and also discuss the effect of different q to the PageRank, as well as how a PageRank will be changed if more links are added to one page or removed from some pages.***

***Web is expending day by day and people generally rely on search engine to explore the web. In such a scenario it is the duty of service provider to provide proper, relevant and quality information to the internet user against their query submitted to the search engine. It is a challenge for service provider to provide proper, relevant and quality information to the internet user by using the web page contents and hyperlink between the web pages. This paper deals with analysis and comparison of web page ranking algorithms based on various parameter to find out their advantages and limitations for the ranking of the web pages. Based on the analysis of different web page ranking algorithms, a comparative study is done to find out their relative strengths and limitations to find out the further scope of research in web page ranking algorithm.***

--------------------------♦----------------------------

## INTRODUCTION

Ranking is an integral component of any information retrieval system. In the case of Web search, because of the size of the Web and the special nature of the Web users, the role of ranking becomes critical. It is common for Web search queries to have thousands or millions of results. On the other hand, Web users do not have the time and patience to go through them to find the ones they are interested in. It has actually been documented that mostWeb users do not look beyond the first page of results. Therefore, it is important for the ranking function to output the desired results within the top few pages, otherwise the search engine is rendered useless.

Search engines are huge power factors on the Web, guiding people to information and services. Google1 is the most successfull search engine in recent years, mostly due to its very comprehensive and accurate search results. When Google was an early research project at Stanford, several papers were written describing the underlying algorithms. The dominant algorithm was called PageRank and is still the key for providing accurate rankings for search results.

Google uses the Power method to compute PageRank. For the whole Internet and larger domains this is probably the only possible method (principally due to the high memoryrequirements of other methods). In the Power method a number (50-100) of matrix vector multiplications are performed.

For smaller domains, other methods than the Power method would be interesting to investigate. One good candidate is the Arnoldi method which has higher memory requirements but converges after less iterations. To efficiently handle these large-scale computations we need to implement the algorithms using a parallel system. Some sort of load balancing might be needed to get good performance for the parallelization.

A Web-crawler needs to be implemented to gather realistic test data. In this review we investigate these methods. As the volume of information on the internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet user. Figure 1 shows a working of a typical search engine, which shows the flow graph for a searched query by a web user.

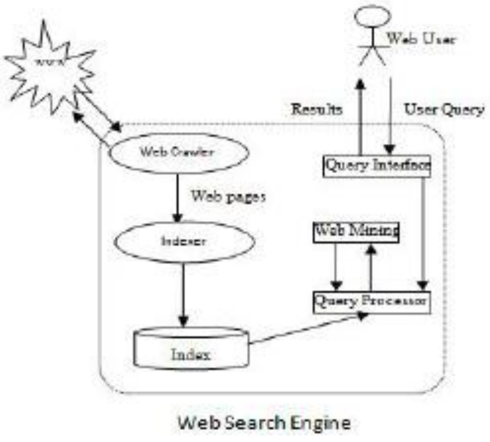

Figure 1: Working of Search Engine

An efficient ranking of query words has a major role in efficient searching for query words. There are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self-descriptiveness. For ranking of web pages, several algorithms are proposed in the literatures.

The motive behind this paper to analyze the currently important algorithms for ranking of web pages to find out their relative strengths, limitations and provide a future direction for the research in the field of efficient algorithm for ranking of the web pages.

## A RANKING FOR EVERY PAGE ON THE WEB

There has been a great deal of work on academic citation analysis. Goffman has published an interesting theory of how information flow in a scientific community is an epidemic process. There has been a fair amount of recent activity on how to exploit the link structure of large hypertext systems such as the web.

Finally, there has been some interest in what "quality" means on the net from a library com- munity. It is obvious to try to apply standard citation analysis techniques to the web's hypertextual citation structure. One can simply think of every link as being like an academic citation. So, a major page like http://www.yahoo.com/ will have tens of thousands of backlinks (or citations) pointing to it.

This fact that the Yahoo home page has so many backlinks generally imply that it is quite important. Indeed, many of the web search engines have used backlink count as a way to try to bias their databases in favor of higher quality or more important pages. However, simple backlink counts have a number of problems on the web. Some of these problems have to do with characteristics of the web which are not present in normal academic citation databases.

**Link Structure of the Web -** While estimates vary, the current graph of the crawl able Web has roughly 150 million nodes (pages) and 1.7 billion edges (links). Every page has some number of forward links (outedges) and backlinks (inedges). We can never know whether we have found all the backlinks of a particular page but if we have downloaded it, we know all of its forward links at that time.

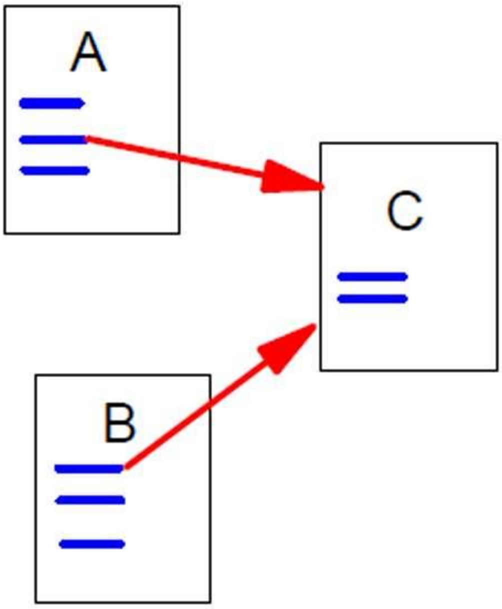

Figure 2: A and B are Backlinks of C

Web pages vary greatly in terms of the number of backlinks they have. For example, the Netscape home page has 62,804 backlinks in our current database compared to most pages which have just a few backlinks. Generally, highly linked pages are more "important" than pages with few links. Simple citation counting has been used to speculate on the future winners of the Nobel Prize. PageRank provides a more sophisticated method for doing citation counting.

The reason that PageRank is interesting is that there are many cases where simple citation counting does not correspond to our common sense notion of importance. For example, if a web page has a link off the Yahoo home page, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places. PageRank is an attempt to see how good an approximation to "importance" can be obtained just from the link structure.

**Propagation of Ranking Through Links -** Based on the discussion above, we give the following intuitive description of PageRank: a page has high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks.

## PAGERANK

In this following section we present the basic ideas of PageRank. We also describe various problems for calculating PageRank and how to resolve them.

PageRank explained - The Internet can be seen as a large graph, where the Web pages themselves represent nodes, and their links (direct connection to other Web pages) can be seen as the edges of the graph. The links (edges) are directed; i.e. a link only points one way, although there is nothing stopping the other page from pointing back. This interpretation of the Web opens many doors when it comes to creating algorithms for deciphering and ranking the world's Web-pages.

The PageRank algorithm is at the heart of the Google search engine. It is this algorithm that in essence decides how important a specific page is and therefore how high it will show up in a search result. The underlying idea for the PageRank algorithm is the following: a page is important, if other important pages link to it. This idea can be seen as a way of calculating the importance of pages by voting for them. Each link is viewed as a vote - a de facto recommendation for the importance of a page - whatever reasons the page has for linking to a speci_c page. The PageRank-algorithm can, with this interpretation, be seen as the counter of an online ballot, where pages vote for the importance of others, and this result is then tallied by PageRank and is reected in the search results.

## PAGE RANK ALGORITHM

Page Rank algorithm is the most commonly used algorithm for ranking the various pages. Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The Page Rank considers the back link in deciding the rank score. If the addition of the all the ranks of the back links is large then the page then it is provided a large rank. A simplified version of PageRank is given by:

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

Where the PageRank value for a web page u is dependent on the PageRank values for each web page v out of the set Bu (this set contains all pages linking to web page u), divided by the number L(v) of links from page v.

An example of back link is shown in figure 3 below. U is the back link of V & W and V & W are the back links of X.

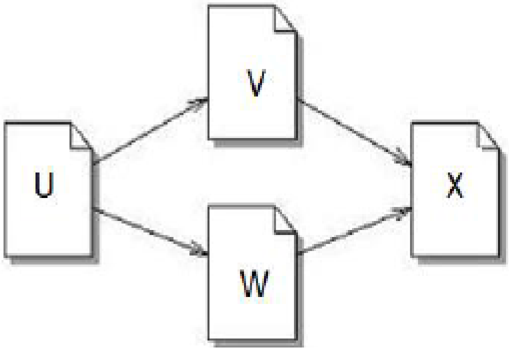

Figure 3: Illustration of back links

**HITS Algorithm -** HITS algorithm ranks the web page by processing in links and out links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyperlinks and a web page is named as HUB if the page point to various hyperlinks. An Illustration of HUB and authority are shown in figure 4.

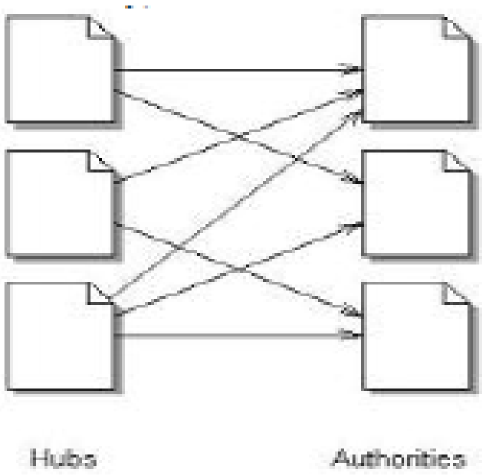

Figure 4: Illustration of Hub and Authorities

HITS is technically, a link based algorithm. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents.

Original HITS algorithm has some problems which are given below.

(i) High rank value is given to some popular website that is not highly relevant to the given query.

(ii) Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the outlinks of a hub page. Figure 5 shows an Illustration of HITS process.

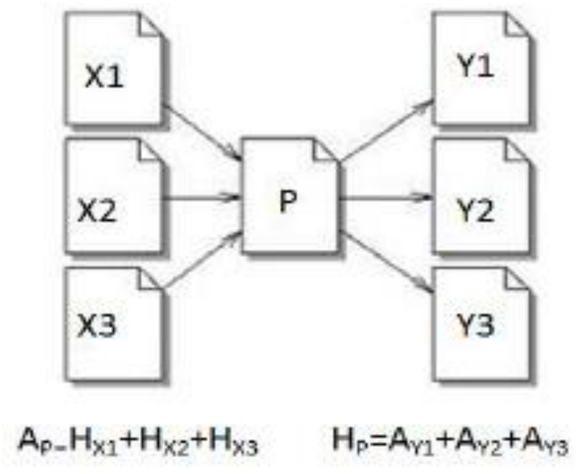

Figure 5: Illustration of HITS process

To minimize the problem of the original HITS algorithm, a clever algorithm is proposed by reference. Clever algorithm is the modification of standard original HITS algorithm. This algorithm provides a weight value to every link depending on the terms of queries and endpoints of the link. An anchor tag is combined to decide the weights to the link and a large hub is broken down into smaller parts so that every hub page is concentrated only on one topic.

**Weighted Page Rank Algorithm -** Weighted Page Rank Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages.

This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among its out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links.

Simulation of WPR is done using the website of Saint Thomas University and simulation results show that WPR algorithm finds larger number of relevant pages compared to standard page rank algorithm. As suggested by the author, the performance of WPR is to be tested by using different websites and future work include to calculate the rank score by utilizing more than one level of reference page list and increasing the number of human user to classify the web pages.

**Weighted Links Rank Algorithm -** A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis named as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are improved using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented. Future work in this algorithm includes, tuning of the weight factor of every term for further evolution.

**EigenRumor Algorithm -** As the number of blogging sites is increasing day by day, there is a challenge for service provider to provide good blogs to the users. Page rank and HITS are very promising in providing the rank value to the blogs but some limitations arise, if these two algorithms are applied directly to the blogs The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, a EigenRumor algorithm is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector.

**Distance Rank Algorithm -** An intelligent ranking algorithm named as distance rank is proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani. It is based on reinforcement learning algorithm. In this algorithm, the distance between pages is considered as a punishment factor. In this algorithm the ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them.

The Advantage of this algorithm is that it can find pages with high quality and more quickly with the use of distance based solution. The Limitation of this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages.

**Time Rank Algorithm -** An algorithm named as TimeRank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al. Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

**TagRank Algorithm -** A novel algorithm named as TagRank for ranking the web page based on social annotations is proposed by Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun. This algorithm calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very accurate and this algorithm index new information resources in a better way. Future work in this direction can be to utilize cooccurrence factor of the tag to determine weight of the tag and this algorithm can also be improved by using semantic relationship among the co-occurrence tags.

**Relation Based Algorithm -** Fabrizio Lamberti, Andrea Sanna and Claudio Demartini proposed a relation based algorithm for the ranking the web page for semantic web search engine. Various search

engines are presented for better information extraction by using relations of the semantic web. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Results are very encouraging on the parameter of time complexity and accuracy. Further improvement in this algorithm can be the increased use of scalability into future semantic web repositories.

## PAGERANK IMPLEMENTATION

We convert, each URL into a unique integer, and store each hyperlink in a database using the integer IDs to identify pages. Details of our implementation are in. In general, we have implemented Page Rank in the; following maimer. First we sort the link structure by Parent II). Then dangling links are removed from the link database for reasons discussed above (a few iterations removes the vast majority of the dangling links). We need to make an initial assignment of the ranks. This assignment can be made by one; of several strategies. If it is going to iterate until convergence, in general the initial values will not affect, final values, just the rate of convergence. But, we can speed up convergence by choosing a good initial assignment. We believe that careful choice of the initial assignment and a small finite number of iterations may result in excellent or improved performance.

Memory is allocated for the; weights for every page. Since we use single precision floating point values at 1 bytes each, this amounts to 300 megabytes for our 75 million URLs. If insufficient RAM is available to hold all the weights, multiple passes can be made (our implementation uses half as much memory and two passes). The weights from the current time; step art! kept in memory, and the previous weights an; accessed linearly on disk. Also, all the access to the link database!, *A,* is linear because it is sorted. Therefore, *A* can be kept on disk as well. Although these data structures an; very large, linear disk access allows each iteration to be completed in about fi minutes on a typical workstation. After the weights have converged, we add the dangling links back in and recompute the rankings. Note after adding the dangling links back in, we need to iterate as many times as was required to remove the dangling links. Otherwise, some of the dangling links will have a zero weight. This whole; process takes about, five hours in the; current implementation. With less strict, convergence criteria, and mem; optimization, the; calculation could be; much faster. Or, more; efficient, techniques for estimating c;igcnvcctors could be; used to improve performance. However, it, should be; noted that, the; cost required to compute the; PageRank is insignificant compared to the cost required to build a full text index.

## CONCLUSION

Finding a high-quality ranking in a sparsely connected graph is a challenging and yet meaningful task in the era of Internet computing and social networks. In this article, we reviewed the famous PageRank algorithm that is widely adopted by most leading search engine companies such as Google and Baidu and provided thorough analysis of its power iteration. We proposed a few methods to improve the quality of the ranking generated by PageRank. Our experiments showed that the results of quality ranking can be greatly improved when the random-page-hop parameter *q* is well-tuned by our proposed strategy. With the help of the proposed strategy, search engines will be able to find more relevant search results for the end users.

Based on the algorithm used, the ranking algorithm provides a definite rank to resultant web pages. A typical search engine should use web page ranking techniques based on the specific needs of the users. After going through exhaustive analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the results, it is concluded that existing techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

## REFERENCES

- Ali Mohammad Zareh Bidoki and Nasser Yazdani, “DistanceRank: An Iintelligent Ranking Algorithm for Web Pages”, Information Processing and Management, 2007.

- C. Ridings and M. Shishigin, “Pagerank Uncovered”, Technical Report, 2002.

- Fabrizio Lamberti, Andrea Sanna and Claudio Demartini , "A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.

- J Xie, XLu, S Shao. *An improvement algorithm research on S-MAC protocol based on adaptive backoff window*, Comput Moder, 2009, 12: 119-121.

- Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

- L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

- Milan Vojnovic et al., "Ranking and Suggesting Popular Items", In IEEE Transaction of KDE, Vol. 21, No. 8, Aug 2009.

- Neelam Duhan, A. K. Sharma and Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey", In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

- Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 33:107{117, 1998.