



# AN ENHANCED PAGE RANKING ALGORITHM USING DYNAMIC APPROACH

Er. Jasdeep Singh <sup>a,\*</sup>, Prince Sharma <sup>b</sup>

<sup>a</sup> Department of CSE, Assistant. Professor

<sup>b</sup> Department of CSE, Student

---

**ABSTRACT:** We live in the computer era. The World Wide Web is part of our daily lives and information is only a click away. As the World Wide Web is increasing regularly, therefore the most concerned issue for the users would be how to collect the valuable information as well as get their genuine information efficiently and instantly. Hence, simply open your favorite search engine, like Google or Yahoo, type in the key words, and the search engine will display the web pages relevant for your search. Nowadays, the ranking of the web pages for the web search engine is one of the significant problems. So, this leads to the important attention to the research community. Web mining performs an important role in extracting the best information according to a user's problem. In this paper, a survey of a new page rank algorithm which is an enhanced page ranking algorithm using dynamic approach. This algorithm is used to find more relevant information according to a user's query.

**KEYWORDS:** WWW; Data mining; Web mining; Search engine; Page ranking

---

## I. INTRODUCTION

The World Wide Web is a really helpful information tool which is used for information retrieval and knowledge discovery. The vast pool of information related to the user queries is stored in different web pages and is growing day by day. However, all information is not relevant to the user. In such situations, different web search engines are used by users for finding appropriate answers for their queries. Every web search engine has its own architecture and based on that, it performs number of tasks. Retrieving relevant information and providing it to users has become increasingly essential. So, ranking algorithms are used to sort the web pages so that more relevant results are displayed at the top. In this Paper, An enhanced page ranking algorithm using dynamic approach is being proposed for search engines that work on the basis of page rank algorithm. The important purpose of this proposed algorithm is to find more relevant information as per the queries of the user.

## WEB MINING

The Page rank plays an important role in the process of web mining. Web mining is a data mining technique in which large numbers of web pages are crawled and knowledge is extracted from them. Web Mining is a Data Mining technique used in discover different patterns from the web. Web mining can be broadly classified into:-

**Web Content Mining:** WCM is the process of extract the useful information and knowledge from the web page content.

**Web Structure Mining:** WSM mainly analyze the node and link structure of a web site.

**Web Usage Mining:** WUM is the process of discovering the data stored in server access logs, user profile and pattern in user browsing the web pages.

### Page Rank Algorithm

Page Rank is an algorithm used by Google Search to rank websites in their search engine results. Page Rank is a way of measuring the importance of website pages. According to Google: Page Rank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

Page Rank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element  $E$  is referred to as the *Page Rank of  $E$*  and denoted by  $PR(E)$ . Other factors like Author Rank can contribute to the importance of an entity.

A Page Rank results from a mathematical algorithm based on the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. 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"). A page that is linked to by many pages with high Page Rank receives a high rank itself.

### The Page Rank Algorithm

The original Page Rank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

where

- $PR(A)$  is the Page Rank of page A,
- $PR(T_i)$  is the Page Rank of pages  $T_i$  which link to page A,
- $C(T_i)$  is the number of outbound links on page  $T_i$  and
- $d$  is a damping factor which can be set between 0 and 1.

So, first of all, we see that Page Rank does not rank web sites as a whole, but is determined for each page individually. Further, the Page Rank of page A is recursively defined by the Page Ranks of those pages which link to page A.

The Page Rank of pages  $T_i$  which link to page A does not influence the Page Rank of page A uniformly. Within the Page Rank algorithm, the Page Rank of a page T is always weighted by the number of outbound links  $C(T)$  on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.

The weighted Page Rank of pages  $T_i$  is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's Page Rank.

Finally, the sum of the weighted Page Ranks of all pages  $T_i$  is multiplied with a damping factor  $d$  which can be set between 0 and 1. Thereby, the extend of Page Rank benefit for a page by another page linking to it is reduced.

## II. PREVIOUS WORK DONE

**Wenpu Xing (2004) [1]** states that with the rapid growth of the Web, users get easily lost in the rich hyper structure. Providing relevant information to the users to cater to their needs is the primary goal of website owners. Therefore, finding the content of the Web and retrieving the users' interests and needs from their behavior have become increasingly important. Web mining is used to categorize users and pages by analyzing the users' behavior, the content of the pages, and the order of the URLs that tend to be accessed in order. Web structure mining plays an important role in this approach. Two page ranking algorithms, HITS and Page Rank, are commonly used in web structure mining. Both algorithms treat all links equally when distributing rank scores. Several algorithms have been developed to improve the performance of these methods. The Weighted Page Rank algorithm (WPR), an extension to the standard Page Rank algorithm, is introduced in this paper. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. The results of simulation studies show that WPR performs better than the conventional Page Rank algorithm in terms of returning larger number of relevant pages to a given query.

**Waiki Ching(2005) [2]** states that the computation of Page Rank is an important issue in Internet. Especially one has to handle a huge size of web with its size growing rapidly. In this paper, we present an adaptive numerical method for solving the Page Rank problem. The numerical method combines the Jacobi Over-Relaxation (JOR) method with the evolutionary algorithm.

**R.Manoharan(2011) [3]** states that web services are dynamic applications that are published over the network to help in data exchange between systems and other applications. The selection of these web services is an important part of Web service composition. As the number of Web services on the internet increase, the need for finding the exact web service that matches the user's request also increases. So ranking of web services is required in order to find the right web service. Earlier methods of ranking involved using a matrix to rank the web services by using their QoS property. But in many cases it might be tedious to define the QoS values accurately. So a fuzzy logic was proposed to deal with the improper QoS constraints. Many approaches have been proposed both in Quality based and fuzzy based ranking. This paper takes the advantages of both the methods and propose a new ranking method which is a hybrid of matrix ranking method and QoS based fuzzy ranking method.

**HemaDubey (2011) [4]** conclude that page ranking is an important component for information retrieval system. It is used to measure the importance and behavior of web pages. We review two approaches for ranking: HITS concept and Page Rank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages. The Page Rank algorithm calculates the rank of individual web page and Hypertext Induced Topic Search (HITS) depends upon the hubs and authority framework. A fast and efficient page ranking mechanism for web retrieval remains as a challenge. This paper proposed a new page rank algorithm which uses a normalization technique based on mean value of page ranks. The proposed scheme reduces the time complexity of the traditional Page Rank algorithm by reducing the number of iterations to reach a convergence point.

**Aditya Pratap Singh (2011) [5]** proposed an efficient method to rank the research papers from various fields of research published in various conferences over the years. This ranking method is based on citation network. The importance of a research paper is captured well by the peer vote, which in this case is the research paper being cited in other research papers. Using a modified version of the Page Rank algorithm, they rank the research papers, assigning each of them an authoritative score. Using the scores of the research papers calculated by above mentioned method, we formulate scores for conferences and authors and rank them as well. A new metric in the algorithm which takes into account the time factor in ranking the research papers to reduce the bias against the recent papers which get less time for being studied and consequently cited by the researchers as compared to the older papers. Often a researcher is more interested in finding the top conferences in a particular

year rather than the overall conference ranking. Considering the year of publication of the papers, in addition to the paper scores we also calculated the year wise score of each conference by slight improvisation of the above mentioned algorithm.

**Christian Borgs (2012) [6]** developed a nearly optimal, sublinear time, randomized algorithm for a close variant of this problem. When given a directed network  $G = (V, E)$ , a threshold value  $\Delta$ , and a positive constant  $c > 3$ , with probability  $1 - o(1)$ , our algorithm will return a subset  $S \subseteq V$  with the property that  $S$  contains all vertices of Page Rank at least  $\Delta$  and no vertex with Page Rank less than  $\Delta/c$ . The running time of our algorithm is always  $O(n \Delta)$ . In addition, this algorithm can be efficiently implemented in various network access models including the Jump and Crawl query model recently studied by making it suitable for dealing with large social and information networks. As part of our analysis, we show that any algorithm for solving this problem must have expected time complexity of  $\Omega(n \Delta)$ . Thus, our algorithm is optimal up to logarithmic factors. Our algorithm (for identifying vertices with significant Page Rank) applies a multi-scale sampling scheme that uses a fast personalized Page Rank estimator as its main subroutine. For that, we develop a new local randomized algorithm for approximating personalized Page Rank which is more robust than the earlier ones developed by Jeh and Widom and by Andersen, Chung, and Lang.

**Ms.N.Preeth (2012) [7]** states that web information retrieval deals with a technique of finding relevant web pages for any given query from a collection of documents. Search engines have become the most helpful tool for obtaining useful information from the Internet. The next generation Web architecture, represented by the Semantic Web, provides the layered architecture possibly allowing data to be reused across application. The proposed architecture use a hybrid methodology named Case and Relation (CARE) based Page Rank algorithm which uses past problem solving experience maintained in the case base to form a best matching relations and then use them for generating graphs and spanning forests to assign a relevant score to the pages.

**Sharmila Subudhi (2013) [8]** states that web is the most important tool in now-a-days upon which people rely on to search their required information. In such a scenario it is the duty of service provider to provide proper, relevant and quality information to the internet where user can submit their query and find out the result. But 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. The next generation Web architecture, represented by the Semantic Web, provides the layered architecture possibly allowing overcoming this limitation. Several search engines have been proposed, which allow increase in information retrieval accuracy by exploiting keywords and their relations. This paper deals with a hybrid approach of page ranking algorithm which simply based on the prediction and calculation of different numbers of back-links to a web page.

**Madhurdeep Kaur, (2014) [9]** proposed a Hybrid Page Rank Algorithm with an Efficient Approach, as the web is escalating day by day, so the most concerned issue for the users would be how to collect the useful information and to find their genuine information effectively and quickly. With the tremendous growth of information available to end users through the web, search engines play a vital role in retrieving and organizing relevant data for various purposes. The ranking of the web pages for the web search engine is one of the significant problems at present. This leads to the important attention to the research community. In this paper, a page rank mechanism called Hybrid Page Rank Algorithm is proposed which is based on both content and link structure of the web pages. This algorithm is used to find more relevant information according to user's query. It also presents the comparison between Sim Rank Algorithm and the Hybrid Page Rank Algorithm.

### III. PROPOSED FRAMEWORK

#### A. Objectives:

- To enhance the existing page ranking system by introducing multiple parameters which are:
  - Content of web
  - Link Rank
  - Keyword Density
  - Keyword Strength
  - H1 & H2 Tags
  - Title & Description meta Tags
- Transforming static data along with the dynamic approach.

Analyzing the performance of the system by implementing the proposed framework and then calculating the results.

#### B. Outline of the Proposed Technique:

The problem with the conventional technique was consideration of only two parameters that are content of the web and the link rank for page ranking and secondly it was a static approach. The need was to propose a technique considering more parameters for page ranking. So, a new algorithm for page ranking has been proposed that utilizes more parameters for ranking. Along with the content of web and link rank, other factors like keyword density, H1 & H2 ranking and the title page ranking are also considered while ranking pages on the web. More are the parameters considered; more efficient will be ranking of pages. The second updating that is used in the proposed technique is that it works on the basis of both the static and dynamic approach to the performance of the proposed technique can be analyzed by verifying the results on the basis of performance parameters. The proposed technique is an efficient technique for ranking pages on the web.



#### IV. CONCLUSION

The page ranking algorithm which is an application of web mining, web mining is used to extract information from user's queries. As huge amount of information is present on the web, the user's spend a lot of time to get the information that is most related to them. This paper will represent a complete study of page rank algorithm. Various parameters will be applied. We will examine an enhanced page ranking system using dynamic approach as compared to previous static approach. In this way, it helps users to get the relevant information about their queries quickly.

#### REFERENCES

- [1] Madhurdeep Kaur, "A Hybrid Page Rank Algorithm: An Efficient Approach", International Journal of Computer Applications (0975 – 8887) Volume 100 – No.16, pp 58-63, August 2014,
- [2] Sharmila Subudhi, "Designing a Hybrid Page Ranking Algorithm for Semantic Web Search Engine", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3, Issue-1, pp 45-48, October 2013.
- [3] Ms.N.Preeth, "Case and Relation (CARE) based Page Rank Algorithm for Semantic Web Search Engines", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1,pp 329-337, May 2012.
- [4] R.Manoharan, "Hybrid Web Services Ranking Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, pp 452-460 , May 2011.
- [5] Waiki Ching, "A Hybrid Numerical Algorithm for Computing Page Rank", Springer, Current Trends in High Performance Computing and Its Applications,pp 257-264, June 2012.
- [6] Hema Dubey, "An Improved Page Rank Algorithm based on Optimized Normalization Technique", (IJCSIT) Vol. 2 (5), pp 2183-2188, May 2011.
- [7] Wenpu Xing, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research, IEEE,June 2004.
- [8] Christian Borgs, "A Sublinear Time Algorithm for Page Rank Computations", Algorithms and Models for the Web Graph. Springer Berlin Heidelberg,pp 41-53, February 2013.
- [9] Aditya Pratap Singh, "An Efficient Algorithm for Ranking Research Papers based on Citation Network", Data Mining and Optimization (DMO), 3rd Conference on. IEEE,2011.