# STUDY OF WEB PAGE RANK ALGORITHMS

By

POONAM RAWAT \*

**DEEPAK PATHAK \*\*** 

\* Student, G.B.P.U.A & T Pantnagar, India. \*\* Student, Uttarakhand Technical University, India.

#### **ABSTRACT**

World Wide Web is growing every day, that people searching for information can be easily lost in the web. People use search engine to find relevant information from the web. So it is very challenging for search engine to find relevant and quality information from the huge collection of hyperlinked data on World Wide Web. This paper gives an introduction of web search engine with its architecture. It also describes web mining techniques and page ranking algorithms in detail. Some very important page ranking algorithms like PageRank algorithm, Weighted Pagerank algorithm and Hyperlink induced topic search (HITS) algorithm are discussed and compared on the basis of various parameters like mining techniques, input parameter, working process, complexity, limitation and search engine using algorithm.

Keywords: HITS, PageRank, Weighted PageRank, Web Mining.

#### INTRODUCTION

WWW is a huge collection of hyperlinked and heterogeneous information resources, for example: text, images, audios, videos and metadata. As the volume of information on the web is increasing day by day, retrieval of relevant information from the web has become challenging because of some key factors like: Web is vast, Web pages are semi structured, and Information on the web tends to be of diversified meaning, degree of the quality of the information extracted from the web [2, 4]. So it is essential to analyze and understand the underlying data structure of the Web for efficient Information Retrieval. Web mining techniques along with other areas like Natural Language Processing (NLP), Database (DB), Information Retrieval (IR), Machine Learning etc. can be used to solve the challenges. Today majority of the users use information retrieval tools like search engine to find relevant information from the collection of data. Desktop search engine are used to find information from the collection of data from the computer (offline basis), while web search engines are used to retrieve relevant information from the internet. Some most commonly used web search engines are: Google, Yahoo, Bing, MSN etc [2]. All these search engines answer trillions of queries every day.

There are three important components in a web search

engine, which are Crawler, Indexer and Ranking mechanism. The Front-end of the search engine contains- search engine interface, query processor, ranking mechanism and in the back-end, it has crawler and indexer. Crawler is also known as a robot, scooter or spider that recursively navigates the web, downloads the documents and identifies new URLs to visit from the previously visited web pages. The downloaded pages are being transferred to an indexing module and then indexer parses the web pages and builds indexes which are based on the keywords in individual pages and store them on index files [4]. When an end user types query keywords in search engine interface, the query processors match query keywords with the index and returns the URLs of the pages. Then, finally, ranking mechanism is done by

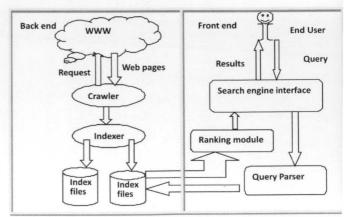


Figure 1. Basic Architecture of Web Search Engine

ranking all these returned pages according to chosen relevancy ranking algorithm. Figure. 1. shows the working of a typical search engine.

This paper is organized as follows. Section 1 describes web mining techniques. In this section web structure mining, web content mining and web usage mining are explained. Section 2 explores different types of Page Rank algorithms like PageRank algorithm [8], Weighted PageRank algorithm [6] and Hypertext Induced selection (HITS) [7]. Section 3 compares all Page Rank algorithms and finally the paper is concluded.

## 1. Web Mining

Data mining is extracting useful information from the huge collection of data. And web mining is an application of data mining, which is used to crawl through various web resources to collect desired information. In other words, web mining is extracting useful information from the web. Web mining can be divided into three categories namely web structure mining, web content mining and web usage mining [9].

### 1.1 Web structure mining

Web structure mining uses the hyperlinked structure of the web documents and generates the structure summary about the web documents in the form of web graph, where web pages are represented as nodes and hyperlinks between web documents which are represented as edges. In this technique, back links and forward links play vital role in the retrieval of the web page.

#### 1.2 Web content mining

In this mining technique, information is extracted from the contents of web documents or result pages produced from search engine. The web document may contain text, multimedia (Audio, video, images) or structured records like lists and tables. Web content mining has two types of approaches: agent based approach and database approaches.

## 1.3 Web usage mining

In this mining technique, user navigation pattern and the useful information is extracted from the secondary data stored in server like access log, agent log, cookies, user profile data, metadata. All the data is derived from the

interactions of the user while surfing on the web.

## 2. Ranking Mechanism

Ranking module plays a vital role in web search engine. It is used by web search engine for ranking search result. For a user is queries, it determines the order of the return pages in result. The order of the web pages basically depends on Page rank of the pages. Web pages' having high Page Rank comes up in the returned result; it means that web pages are arranged in descending order of their Page Rank in return result for a query by search engine. For this purpose, search engine uses page ranking algorithms. Basically page ranking algorithm are used to present the search result by considering the relevance, importance and content score [2]. Some page ranking algorithms depend upon the hyperlinked structure of the web pages which is a type of web structure mining, and some other page ranking algorithms depend only on the actual content on the web pages known as web content mining, while some page ranking algorithms use the combination of hyperlinked structure and content in the web pages [4]. Some of the most popular page ranking algorithms are discussed below.

## 2.1 PageRank Algorithm

The PageRank algorithm was developed by Larry Page and Sergey Brin in 1996 at Stanford University. Page Rank algorithm is based on link analysis which is used by the Google web search engine [8]. Google Search engine is very popular and successful because of its PageRank algorithm. PageRank is a method of measuring a page's "importance." PageRank is a numeric value and it provides a better approach that can compute the importance of the web page by counting the number of web pages that are backlinked. For example: if we have a web page A which links to web page B, then web page A is saying that Web page B is an important page. Pagerank algorithm states that if a web page has some important backlinks to it, then its outgoing links to other web pages also become important [3]. It means backlinks from the important pages play a vital role in the Page Rank of that page (which will be automatically high because of its important backlinks).

Let x be a web page and F(x) is the set of pages that x points to and B(x) is the set of pages that points to page x. Let Nx be the number of the links from page x(outgoing links) and c is a factor of normalization[1]. A simplified version of Page Rank is in "(1)"

$$PR(x) = \sum_{y=B(x)} \frac{PR(y)}{N_y}$$
 (1)

This formula of PageRank is recursive. To compute the PageRank of a page we need to know the PageRank of others. Therefore calculating the PageRanks of pages we started with any random value of PageRank (normally with 1 for all pages in hyperlink structure) and iteratively update these ranks using the above formula and waiting until they converge happens [5]. But there is a small problem with this simplified ranking function. Suppose we have two web pages that point to each other but there are no outgoing links from this loop and there are some web pages which have incoming links to one of them, then during iteration, this loop will accumulate rank but will never distribute any rank, since there are no outgoing links. The loop forms a sort of trap which is known as rank sink [1]. To overcome this problem a modification was made on PageRank formula. Modified Page Rank formula is:

$$PR(x) = (1 - d) + d \sum_{y=B(x)} \frac{PR(y)}{N_y}$$
 (2)

Where d is a damping factor, its value can be set between 0 and 1, we usually set the value of d to 0.85. The damping factor is used to stop the other pages having too much influence because their contribution value on PageRank is damped down by multiplying it by 0.85.

## 2.2 Weighted Page Rank Algorithm (WPR)

Weighted PageRank Algorithm is proposed by Wenpu Xing and Ali Ghorbani. WPR is the modification of the standard PageRank algorithm. WPR does not divide the page rank value of a page equally among its all outgoing or forward linked pages. But it assigns larger page rank values to more important pages than others. And the importance of a web page is measured by its number of backlinks and forward links [6]. The importance is assigned in terms of weighted values of the back links Win(m, n) and forward links Wout(m, n).

$$W^{in}(m,n) = \frac{I_n}{\sum_{P \in R_S(m)} I_P}$$
(3)

$$W^{out}(m,n) = \frac{0_n}{\sum_{p \in Re(m)} o_p}$$
 on (3) is the weight of link(m, n)

calculated based on the number of incoming or back links of page n and the number of incoming or back links of all the reference pages of page m. Ip and In are the number of incoming links of page n and page p respectively. R(m) denotes the reference page list of page m. W<sup>n</sup>(m, n) is as shown in equation (4) which is the weight of Link(m, n) calculated based on the number of outgoing or forward links of page n and the number of outgoing or forward links of all reference pages of m. Where On is the number of outgoing links of page n and Op is the number of outgoing links of page p. WPR formula is shown in equation (5),

$$WPR(n) = (1 - d) + d \sum_{m \in B(n)} WPR(m) W^{in}(m, n) W^{out}(m, n)$$
 (5)

Where d is damping factor, its value can be set between 0 to 1. And B(n) is the set of pages that point to page n. We have four categories of resultant page for a query. Where categorization is based on the resultant page's relevancy to the given query [2].

- Very Relevant Page (VR): This category pages contain very related information to the given query.
- Relevant Page (IR): This category of pages contains relevant information (not important) about the given query.
- Weak Relevant Page (WR): This category of pages may have the query keyword but these pages do not contain any relevant information.
- Irrelevant Page (IR): Irrelevant pages neither contain relevant information nor query keywords.

#### 2.3 Hyperlink Induced Topic Search (HITS)

In 1988 Kleinberg proposed HITS algorithm [7]. He identified two different web pages called hubs and authority. Authority pages are popular and are the ones having relevant information for a given query, while hubs are web pages that contain important links to relevant

pages or many authorities. So, a good authority page is pointed by many good hub pages on the same subject, and good hub pages are having links to good authority pages. A web page may be a good hub and authority at the same time. Figure. 2 is illustrating hub and authority relationship. The HITS algorithm treats hyperlinked structure of WWW as a directed graph G (V, E), where V is a set of vertices representing web pages and E is the set of edges that representing links between pages. The HITS algorithm works in two major steps. First step is sampling and second step is Iterative. In sampling step Hub and Authority weight is calculated using the following equations (6 and 7)".

$$H_p = \sum_{q=l(p)} A_q \tag{6}$$

$$A_p = \sum_{q=B(p)} H_q$$
 (7)

Where  $H_p$  is the hub weight and Ap is authority weight. I(p) and B(p) are the set of the reference and referrer pages of the web page p. In iterative step, the output of the sampling step is used to find Hubs and Authorities. The page's authority weight  $(A_p)$  is proportional to the sum of the hub weights  $(H_{q1}, H_{q2}$  etc) of pages that it points to it, Kleinberg [7]. Similarly, a page's hub weight  $(H_q)$  is proportional to the sum of the authority weights  $(A_{p1}, A_{p2}, A_{p3})$  of pages that it points to. HITS algorithm is used by 'Clever' a prototype search engine for IBM research project'. But HITS could not be implemented in real time search engine because of the topic drift and some efficiency issues [7].

#### 3. A Comparison Study

Based on the analysis, a comparison of the PageRank algorithm, WPR and HITS algorithm is shown in Table 1.

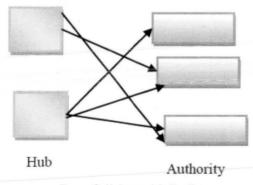


Figure 2. Hubs and Authority

Algorithm Criteria	PageRank	Weighted PageRa	nk HITS	
Mining Technique  •	es Web Structure Mining	Web Structure Mining	Web Structure Mining and Web Content Mining	
Input Parameters	Back links	Back Links and Forward Links	Back Links, Forward Links and content	
Working Process	Computes Page rank value at the index time and result are sorted on the bases of the priority of pages.	Computes Page rank value at the index time and result are sorted on the bases of the importance of pages.	N highly relevant pages are computed and find value on the fly.	
Complexity	O(log N)	< O(log N)	< O(log N)	
Search engine	Google :	Research models	Clever	
Quality of Result	Medium	Higher than PageRank	Less than PageRank	
Lamination	Query Independent, it means Results come at the time of indexing and not at the query time.	Query independent and Relevancy is ignored	Importance of the pages is totally ignoredTopic problem and efficiency issues	

Table 1. Comparison between Ranking Algorithms

Comparison is done on the basis of various parameters like mining techniques, input parameter, working process, complexity, limitation and search engine using algorithm. PageRank and HITS are the most important algorithms among all the algorithms. PageRank algorithm is used by Google Search engine and HITS algorithm is used in the prototype search engine 'Clever' of IBM.

#### Conclusion

The Page ranking Algorithm plays a vital role in the ranking mechanism of search engine. It determines the order of the return result for a query. This paper describes that PageRank algorithm and WPR only depends on the hyperlinked structure of the web, while HITS algorithm stresses on the content of the page as well as the hyperlinked structure of the web. A typical search engine may deploy a particular Page Ranking algorithm according to the user needs. In order to improve the search result, a ranking algorithm should be developed by equally considering the importance and relevancy of a page.

#### References

[1]. 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.

- [2]. Ashutosh Kumar Singh, Ravi Kumar P, (2009). "A Comparative Study of Page Ranking Algorithms for Information Retrieval", International Journal of Electrical and Computer Engineering 4:7.
- [3]. Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia (2009). "Page Ranking Algorithms: A Survey", IEEE International Advance Computing Conference, 978-1-4244-1888-6/08
- [4]. Laxmi Choudhary and Bhawani Shankar Burdak, "Role of Ranking Algorithms for Information Retrieval".
- [5]. Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar, (2013). "Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 978-0-7695-4958-3/13 DOI

- 10.1109/CSNT.137/IEEE.
- [6]. Wenpu Xing and Ali Ghorbani, (2004). "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314,
- [7]. J. Kleinberg, (1999). "Authoritative Sources in a Hyper-Linked Environment", *Journal of the ACM* 46(5), pp. 604-632.
- [8]. S. Brin, and L. Page, (1998). "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117,
- [9]. Cooley, R, Mobasher, B., Srivastava, J. (1997). "Web Mining: Information and pattern discovery on the World Wide Web". In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI' 97). Newposrt Beach, CA.

#### **ABOUT THE AUTHORS**

Er. Poonam Rawat is currently pursuing M-Tech in Information Technology from Govind Ballabh Pant University of Agriculture & Technology, Pantnagar. She has obtained her B-Tech degree in Computer Science Engineering from Dev Bhoomi Institute of Technology (DBIT), Dehradun (U.K.) India in 2012. Her research interest includes Web Search Algorithms and Web Mining.



Er. Deepak Pathak is currently pursuing M-Tech in Information Technology from Uttrakhand Technical University, Dehradun. He has received his B-Tech degree in Computer Science Engineering from Dev Bhoomi Institute of Technology (DBIT), Dehradun (U.K.) India in 2012, His research interest includes Wireless Sensor Network, Computer Networks and Cryptography.



# Price List/Subscription/Renewal Form-2014

Yes, I would like to Subscribe/Renew i-manager's Journals

		Print		Online		Print & Online	
Er	ngineering Journals	Annual	Three Yrs	Annual	Three Yrs	Annual	Three Yrs
	Journal on Future Engineeri & Technology (JFET)	ng	₹ 4800/-	₹ 1400/-	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Software Engineering (JSE)	<b> </b>	₹ 4800/-	₹1400/-	<b> ₹ 3800/-</b>	₹ 2800/-	_ ₹7500/-
	Journal on Electrical Engineering (JEE)	<b> </b>	<b> ₹ 4800/-</b>	_ ₹1400/-	₹ 3800/-	₹ 2800/-	<b>₹ 7500/-</b>
	Journal on Electronics Engineering (JELE)	[ ₹ 1800/-	₹ 4800/-	[ ] ₹ 1400/-	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Mechanical Engineering (JME)	<b> </b>	<b> ₹ 4800/-</b>	☐ ₹ 1400/-	₹ 3800/-	₹ 2800/-	₹ 7500/-
	Journal on Civil Engineering (JCE)	<b> </b>	_ ₹ 4800/-	☐ ₹ 1400/-	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Embedded Systems (JES)	<b> </b>	<b> ₹ 4800/-</b>	<b> </b>	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Information Technology (JIT)	₹ 1800/-	_ ₹4800/-	<u></u> ₹ 1400/-	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Structural Engineering (JSTE)	[ ₹ 1800/-	<b> ₹ 4800/-</b>	<b> </b>	<b>₹ 3800/-</b>	<b> ₹ 2800/-</b>	_ ₹ 7500/-
	Journal on Communication Engineering & Systems (JCS)	₹ 1800/-	<b> </b>	_ ₹1400/-	₹ 3800/-	<b> ₹ 2800/-</b>	_ ₹ 7500/-
	Journal on Wireless Communication Networks (.	₹ 1800/- JWCN)	_ ₹4800/-	₹ 1400/-	₹ 3800/-	₹ 2800/-	_ ₹7500/-
	Journal on Instrumentation Control Engineering (JIC)	&	<b>₹ 4800/-</b>	<b>₹ 1400/-</b>	₹ 3800/-	<b>₹ 2800/</b> -	<b>₹ 7500/-</b>
	Journal on Circuits & System	ns (JCIR)	<b>₹</b> 4800/-	₹ 1400/-	<b>₹ 3800/-</b>	₹ 2800/-	<b>₹ 7500/-</b>
	Journal on Digital Signal Pro (JDP)	ocessing	<b>₹ 4800/</b> -	₹1400/-	₹ 3800/-	₹ 2800/-	₹ 7500/-
	Journal on Power Systems Engineering (JPS)	[ ₹ 1800/-	<b>₹ 4800/</b> -	<b>₹</b> 1400/-	₹ 3800/-	₹ 2800/-	_ ₹ 7500/-
	Journal on Computer Scienc	ee(JCOM)	<b> ₹ 4800/-</b>	₹1400/-	<b> ₹ 3800/-</b>	₹ 2800/-	<b>₹ 7500/-</b>
	Journal on Material Science	(JMS)	<b> </b>	<b> </b>	₹ 3800/-	<b>₹</b> 2800/-	<b>₹ 7500/-</b>
	Journal on Mathematics (JM	1AT) ₹ 1200/-	<b> ₹ 3200/-</b>	₹ 1000/-	₹ 2800/-	<b>₹</b> 2000/-	₹ 5500/-

Journal on Cloud Computing (JCC)	<b>₹ 1800/-</b>	<b>₹ 4800/-</b>	<b>₹ 1400/-</b>	<b>₹ 3800/-</b>	₹ 2800/-	₹ 7500/-
Journal on Mobile Applications & Technologies (JMT)	<b> ₹ 1800/-</b>	<b>₹ 4800/</b> -	_ ₹ 1400/-	<b>₹ 3800/-</b>	₹ 2800/-	₹ 7500/-
Journal on Image Processing (JIP)	<b>₹ 1800/</b> -	☐ ₹4800/-	_ ₹ 1400/-	₹3800/-	_ ₹2800/-	<b>₹ 7500/-</b>
Journal on Pattern Recognition (JPR)	_ ₹ 1800/-	[ ] ₹ 4800/-	₹ 1400/-	₹ 3800/-	<b>₹ 2800/</b> -	<b>₹ 7500/-</b>
Educational Journals						
Journal of Educational Technology (JET)	₹ 1800/-	<b>₹ 4800/</b> -	₹ 1400/-	₹ 3800/-	₹ 2800/-	₹ 7500/-
Journal on School Educational Technology (JSCH)	_ ₹ 1200/-	<b>₹ 3200/</b> -	_ ₹ 1000/-	₹ 2800/-	₹ 2000/-	₹ 5500/-
Journal on Educational Psychology (JPSY)	<b>₹ 1800/-</b>	<b> </b>	₹ 1400/-	₹ 3800/-	₹ 2800/-	₹ 7500/-
Journal on English Language Teaching (JELT)	_ ₹1200/-	<b>₹ 3200/</b> -	₹ 1000/-	₹ 2800/-	_ ₹ 2000/-	₹ 5500/-
Management Journal						
Journal on Management (JMGT)	_ ₹1500/-	<b>₹ 4000/</b> -	<b>₹ 1200/-</b>	_ ₹ 3600/-	_ ₹ 2500/-	₹ 7200/-
Scientific Journals						
Journal on Nursing (JNUR)	₹ 2000/-	_ ₹5200/-	₹ 1500/-	₹ 4500/-	₹ 3000/-	₹ 9000/-
Journal on Pharmacy (JPHARM)	<b>₹ 3000/-</b>	<b>₹ 8000/</b> -	₹2500/-	₹ 7000/-	₹ 5000/-	₹ 14200/-
Journal on Dentistry (JDENTAL)	₹ 3000/-	<b>₹ 8000/</b> -	₹ 2500/-	₹ 7000/-	₹ 5000/-	₹ 14200/-

# **Subscription Form/Renewal Form**

From (Personal Name	):	Position :	
Institution/ Organisati	on's Name/Address:		
E-mail Address :		Website (url) :	
Tel. :	Fax :	Mobile :	
Payment Details :			
We are enclosing a D.	D. For Rs :		
in favour of i-manager	Publications payable at Nage	rcoil.	
D.D. No. :	Date :	Bank Name :	* /

(or)

## Our Bank Account Details:

: i-manager Publications

Account No

: 133602000002000

Bank

: Indian Overseas Bank

Branch

Name

: Meenakshipuram

IFS Code

: IOBA0001336

# Please fill in this order form and mail to:

i-manager Publications,

3/343, Hill View, Town Railway Nagar, Nagercoil,

Kanyakumari Dist., Tamil Nadu-629 001

Ph: 04652-276675

web: www.imanagerpublications.com

