ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Comparative Study of Page Rank and Weighted Page Rank Algorithm

Taruna Kumari[1], Ashlesha Gupta[2], Ashutosh Dixit[3]

M.Tech Student, Department of CE, YMCA University of Science & Technology,Faridabad, India [1]

Assistant Professor, Department of CE, YMCA University of Science & Technology, Faridabad, India [2]

Associate Professor, Department of CE, YMCA University of Science & Technology, Faridabad, India [3]

ABSTRACT: World Wide Web is a distributed heterogeneous information resource which includes data and hyperlinks. With the exponential growth of WWW, it has become difficult to access desired information that matches with user needs and interest. Therefore majority of users today use Search Engine to assist in information retrieval over the Internet. The results retrieved, organized & presented by search engine results in hundreds and millions of linked pages of which many might not be useful to the user. Web page ranking algorithms play an important role in ranking web pages so that the user could retrieve the page which is most relevant to the user's query. Some page ranking algorithms are HITS, PageRank and weighted Pagerank. In this paper, we compare two popular web page ranking algorithms namely: Weighted PageRank algorithm and PageRank algorithm. The paper highlights their variations, strengths, weaknesses and carefully analyzes both algorithms using simulations developed for them.

Keywords: Search Engine, Link based search, PageRank and Weighted PageRank.

I. Introduction

WWW is a vast resource of hyperlinked and heterogeneous information including text, image, audio, video and metadata. From early 1990's WWW has seen an explosive growth. It is estimated that WWW has expanded by about 2000% since its evolution and is doubling in size every six to ten months [1]. With huge increase in availability of information through WWW, it has become difficult to access desired information on Internet; therefore many users use Information retrieval tools like Search Engines to search desired information on the Internet A Search Engine is an information retrieval system which helps users finds information on WWW by making the web pages related to their query available. With a search engine, users have to type in "keywords" relating to the information that they need. The search engine would then return a set of results that match best with the keywords entered. A Web Search Engine can therefore be defined as a software program that takes input from the user, searches its database and returns a set of results. It is important to note that the search engine does not search the internet; rather it searches its database, which is populated with data from the internet by its crawler(s). Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler which follows every link it sees. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed. Data about web pages are stored in an index database for use in later queries. The typical architecture of a search engine is shown in Fig[1].

The major components of search engine are Crawler, Indexer and Query processor. A crawler traverses the web by following hyperlinks and storing downloaded pages in a large database. It starts with seed URL and collects documents by recursively fetching links and storing the extracted URL's into a local repository. The Indexer processes and indexes the pages collected by the crawler. It extracts keywords from each page and records the URL where each word has occurred.

The query engine is responsible for receiving and filling search requests from user. When a user fires a query, query engine receives it and after matching the query keywords with the index, returns the URL's of the pages to the user.

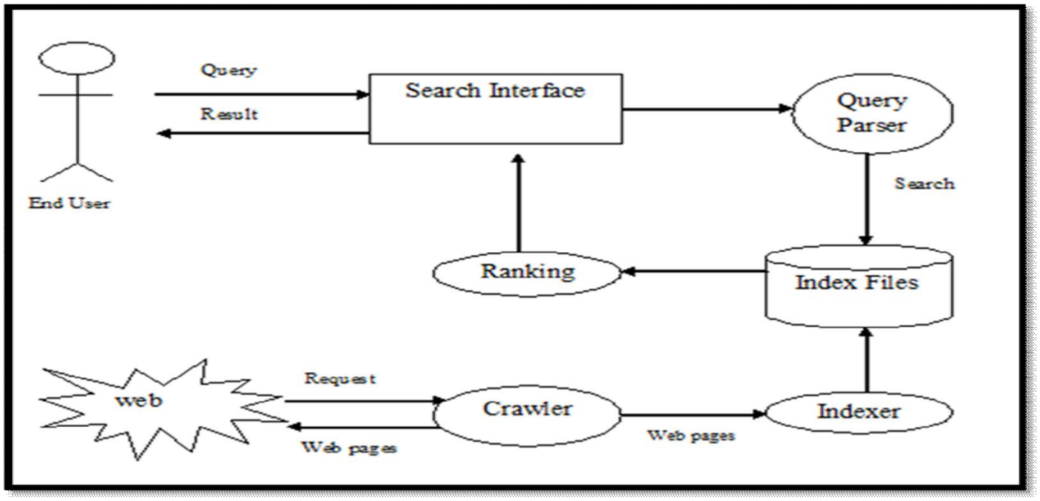


Fig 1. Architecture of a search engine

In general Query Engine may return several hundreds or thousands of URL that match the keywords for a given query. But often users look at top ten results that can be seen without scrolling. Users seldom look at results coming after first search result page, which means that results which are not among top ten are nearly invisible for general user. Therefore to provide better search result, page ranking mechanisms are used by most search engines for putting the important pages on top leaving the less important pages in the bottom of result list. Some of the common page ranking algorithms are PageRank Algorithm [2, 3], Weighted PageRank Algorithm [4] and Hyperlinked Induced Topic search Algorithm [5].

The aim of the paper is to analyse the two popular web page ranking algorithms- Weighted PageRank algorithm and PageRank algorithm and to provide a comparative study of both and to highlight their relative strengths and limitations.

II. PAGE RANKING ALGORITHMS

To present the documents in an ordered manner, Page Ranking methods are applied, which can arrange the documents in order of their relevance, importance and content score. Search engines use two different kinds of ranking factors: query-dependent factors and query Independent Factors .Query-dependent are all ranking factors that are specific to a given query, while query-independent factors are attached to the documents, regardless of a given query. Query-dependent factors used by search engines are measures such as word documents frequency, the position of the query terms within the document or the inverted document frequency, which are all measures that are used in traditional Information Retrieval. Some of the query independent factors are Link popularity, Click popularity and uptodateness etc. Ranking algorithms based on link popularity, falls under Link based ranking algorithm category.

III. LINK BASED PAGE RANKING ALGORITHMS

Link based Page ranking algorithms are based on link structure of the web document. They view the web as a directed graph where the web pages form the nodes and the hyperlinks between the web pages form the directed edges between these nodes [6]. Two important Link Based Page algorithms are given below :

- PageRank[2]
- Weighted PageRank [5].

A. PageRank

Brin and Page [2] developed a ranking algorithm at Stanford University named PageRank after Larry Page. Page Rank algorithm uses link structure to determine the importance of web page. This algorithm is based on random surfer model. The random surfer model assumes that a user randomly keeps on clicking the links on a page and if she/he get bored of a page then switches to another page randomly. Thus, a user under this model shows no bias towards any page or link. PageRank(PR) is the probability of a page being visited by such user under this model.

Page Rank algorithm assumes that if a page has a link to another page then it votes for that page. Therefore, each inlink to a page raises its importance. PageRank is a recursive algorithm in which the PageRank of a page depends upon the PageRank of the pages linking to it. Thus, not only the number of inlinks of a page influences its ranking but also the page ranks of the pages linking to it. A page confers importance to the pages it references to by evenly distributing its PageRank value among all it's outlinks.

The PageRank of page P is given as, follows:

$$PR(P) = 1 - d \ + d \ \sum_{i=0}^{n} \frac{PR(Ni)}{O(Ni)}$$

Where

$N_{0..}.N_n$ are the pages that point to page P.

$O(N_i)$ is defined as the number of links going out of page P.

The parameter d is a Damping factor which can be set between 0 and 1.

Damping factor, d is the probability of user's following the direct links and 1- d denotes the rank distribution from non – directly linked web pages. It is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points. As is evident from the above equations, even if a page doesn't have any inlinks it still has a minimum PR value of 1-d.

Following is a simplified example of the PR algorithm. Consider web graph shown in fig3.

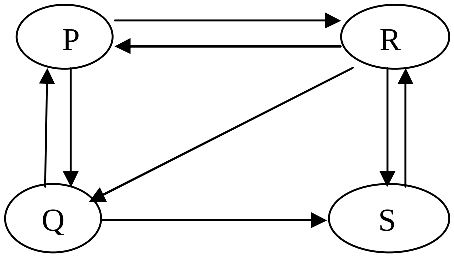


Fig 3

Page, P being referenced by pages Q and R .Q, R has 2,3 outlinks. Then pageRank value of the page P is given as:

$$PR(P)=1-d + d(PR(Q)/2 + PR(R)/3)$$

Iterative Method of Page Rank

In iterative calculation, each page is assigned a starting page rank value of 1. These rank values are iteratively substituted in page rank equations to find the final values. In general, much iteration could be followed to normalize the page ranks.

The PageRank algorithm can be iteratively applied as:

1) Initially let Page rank of all web pages is one
2) Calculate page ranks of all pages by using above formula.
3) Repeat step 2 until values of two consecutive iterations match.

1) Advantages:

- Less Time consuming:- As pageRank is a query independent algorithm i.e. it precomputes the rank score so it takes very less time .
- Feasibility:-This algorithm is more feasible as it computes rank score at indexing time not at query time.
- Importance: - It returns important pages as Rank is calculated on the basis of the popularity of a page.
- Less susceptibility to localized links: - For calculating rank value of a page, it consider the entire web graph, rather than a small subset, it is less susceptible to localized link spam.

2) Disadvantages:

- The main disadvantage is that it favours older pages, because a new page, even a very good one, will not have many links unless it is part of an existing web site.
- Relevancy of the resultant pages to the user query is very less as it does not consider the content of web page.
- Dangling link: This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link.
- Rank Sinks: The Rank sinks problem occurs when in a network pages get in infinite link cycles.
- Dead Ends: Dead Ends are simply pages with no outgoing links.
- Spider Traps: Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.
- Circular References**:** If you have circle references in your website, then it will reduce your front page’s PageRank.

A. Weighted PageRank

Wenpu Xing and Ali Ghorbani [5] projected a Weighted PageRank (WPR) algorithm which is a modification to the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}(u,v)$ and $W^{out}(u,v)$ respectively.

$W^{in}(u,v)$is the weight of link (u, v) calculated based on the number of incoming links of page v and the number of incoming links of all orientations pages of page u.

$$W^{in}(u,v) = \frac{I_v}{\sum_{p\epsilon R(u)} I_p}$$

Where Iv and Ip are the number of incoming links of page v and page p respectively. R (u) denotes the allusion page list of page u.

$$W^{out}(u,v) = \frac{O_v}{\sum_{p\epsilon\ R(u)} O_p}$$

$W^{out}(u,v)$ is the weight of link (u,v) calculated based on the number of outgoing links of page v and the number of outgoing links of all orientations pages of page u.

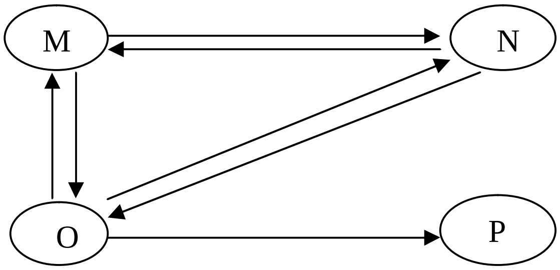


Fig 4.

The weights of incoming as well as outgoing links can be calculated as:

$$W^{in}(n,m) = I_m/(I_m+I_o) = 2/(2+2)=2/4$$

$$W^{out}(n,m) = O_m/(O_m+O_o)=1/(1+3)=1/4$$

The formula as proposed by Wenpu et al for the WPR which is a modification of the PageRank formula is given as

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

1) *Advantages*:
   - Quality: Quality of the pages returned by this algorithm is high as compared to pageRank algorithm.
   - Efficiency: It is more efficient than pageRank because rank value of a page is divided among it's outlink pages according to importance of that page.
2) *Disadvantages*:
   - Less Relevant: As this algorithm considers only link structure not the content of the page, it returns less relevant pages to the user query.

IV. EXPERIMENTS

The WPR and pageRank algorithms were implemented and their results are being compared. Fig 5 illustrates different components involved in the implementation and evaluation of these algorithms. The simulation studies carried out in this work consist of following activities:

1. Finding a web site: Finding a web site with rich hyperlinks is necessary because the standard PageRank and the WPR algorithms rely on the web structure. Insert these pages into database.
2. Extracting URL : This module will extract links within the given page i.e outlinks of a web page. Then these outlink details are also added in the database.

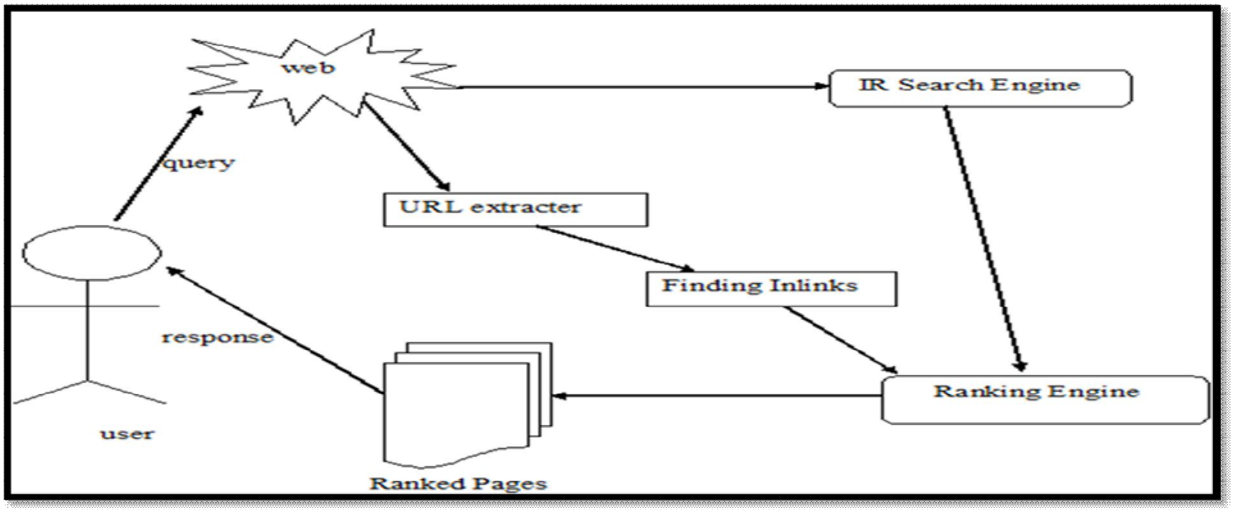


Fig 5. Architectural components of the system used for implementation and evaluation

3. Finding Inlinks: From these extracted outlinks, inlinks can be found and stored in database.
4. Applying algorithms: The Standard PageRank and the WPR algorithms are applied to the pages stored in the database.
5. Evaluating the results: The algorithms are evaluated by comparing their results.

V. EVALUATION

A website that contains many web pages is considered. These web pages can be represented as a web graph as shown below in fig 6.

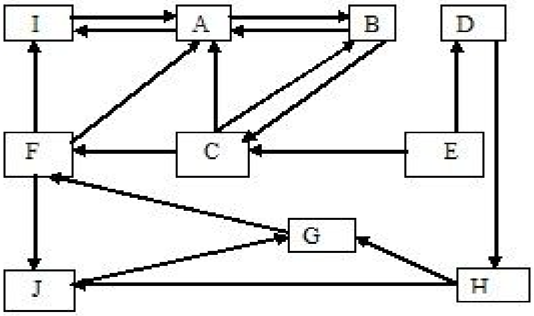
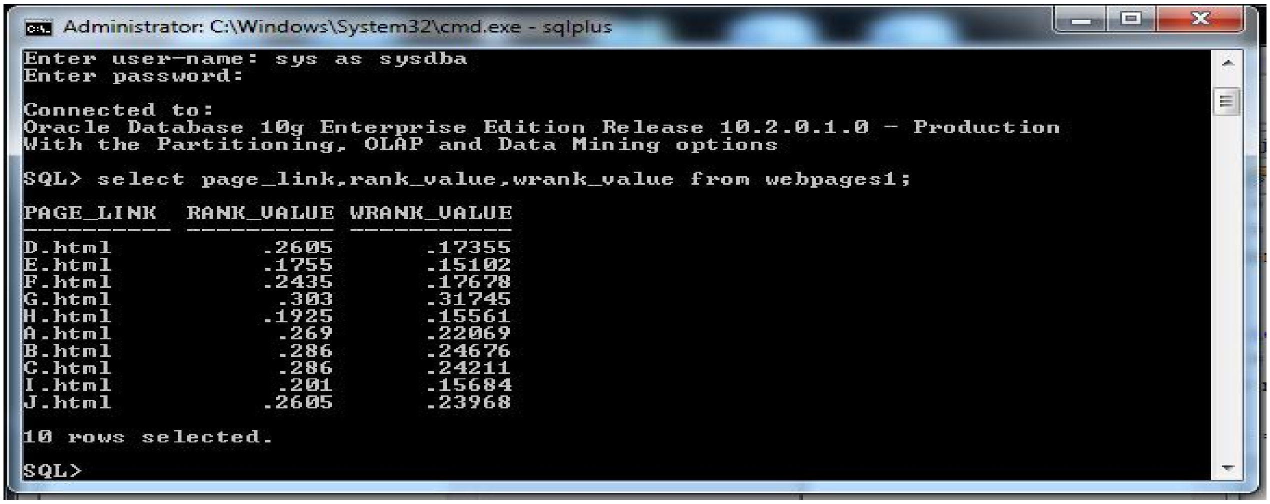


Fig 6. Example figure for analysis

In this graph, nodes(A,B,…,J) are representing web pages pageA , pageB,……., pageJ and links between the web pages are represented by edges . Rank Score of web pages according to pageRank algorithm, rank_value and rank score of web pages by weighted pageRank algorithm , wrank_value are shown below in the fig 7.



Administrator: C:\Windows\System32\cmd.exe - sqlplus

```
Enter user-name: sys as sysdba
Enter password:

Connected to:
Oracle Database 10g Enterprise Edition Release 10.2.0.1.0 - Production
With the Partitioning, OLAP and Data Mining options

SQL> select page_link,rank_value,wrank_value from webpages1;

PAGE_LINK  RANK_VALUE WRANK_VALUE
---------- ---------- -----------
D.html          .2605      .17355
E.html          .1755      .15102
F.html          .2435      .17678
G.html           .303      .31745
H.html          .1925      .15561
A.html           .269      .22069
B.html           .286      .24676
C.html           .286      .24211
I.html           .201      .15684
J.html          .2605      .23968

10 rows selected.

SQL>
```

Fig 7. rank and wrank values of web pages at damping factor (d) = 0.85

As both of these algorithms work iteratively. The rank_value and wrank_value of web pages in various iterations is shown in below table I.

TABLE I
RANK AND WRANK VALUES OF PAGES IN VARIOUS ITERATIONS

	Iterations					
	1		**2**		**3**	
	Rank	**Wrank**	**Rank**	**Wrank**	**Rank**	**Wrank**
pageA	0.269	0.22069	0.4475	0.291	0.609	0.32765
pageB	0.286	0.24676	0.49	0.333	0.634	0.36665
pageC	0.286	0.24211	0.4475	0.3312	0.575	0.36518
PageD	0.2605	0.17355	0.3625	0.1877	0.4305	0.19
PageE	0.1755	0.15102	0.2095	0.1522	0.2435	0.15297
pageF	0.2435	0.17678	0.405	0.2159	0.5325	0.23737
pageG	0.303	0.31745	0.507	0.4443	0.66	0.47302
pageH	0.1925	0.15561	0.2605	0.1597	0.303	0.16053
pageI	0.201	0.15684	0.269	0.1660	0.345	0.1714
pageJ	0.2605	0.23968	0.3795	0.2905	0.49	0.2958

Effect of damping factor

The probability for the random surfer not stopping to click on links is given by the damping factor d, which depends on probability therefore, is set between 0 and 1. The higher d is, the more likely will the random surfer keep clicking links. Since the surfer jumps to another page at random after he stopped clicking links, the probability therefore is implemented as a constant (1-d) into the algorithm. Effect of this constant value is shown in table II.

TABLE II
RANK AND WRANK VALUES OF PAGES AT VARIOUS DAMPING FACTORS

	Damping factor (d)					
	0.15		0.5		0.85	
	Rank	**Wrank**	**Rank**	**Wrank**	**Rank**	**Wrank**
pageA	0.871	0.8624	0.57	0.5415	0.269	0.22069
pageB	0.874	0.8670	0.58	0.5569	0.286	0.24676
pageC	0.874	0.8664	0.58	0.5541	0.286	0.24211
paged	0.8695	0.8541	0.565	0.5138	0.2605	0.17355
Page	0.8545	0.8501	0.515	0.5006	0.1755	0.15102
pageF	0.8665	0.8547	0.555	0.5157	0.2435	0.17678
pageG	0.877	0.8795	0.59	0.5987	0.303	0.31745
pageH	0.8575	0.8509	0.525	0.5033	0.1925	0.15561
pageI	0.859	0.8512	0.53	0.5040	0.201	0.15684
pageJ	0.8695	0.8658	0.565	0.5527	0.2605	0.23968

VI. COMPARISON OF PAGERANK AND WEIGHTED PAGERANK ALGORITHMS

Table III below enlists the comparison of pageRank and weighted pageRank algorithm.

Table III
COMPARISON OF WEIGHTED PAGE RANK AND PAGERANK ALGORITHM

Criteria	**pageRank**	**Weighted pageRank**
Basic criteria	graph based ranking algorithm. Consider only back link in rank calculation.	Based on the calculation of weight of the page with consideration of incoming and outgoing links.
Technique used	Web structure mining	Web structure mining
Description	Compute rank score at indexing time by equally dividing rank value of a page among its outlink pages.	Compute rank score at indexing time by unequally dividing rank value of a page among its outlink pages.
Complexity	O(Log(N*))	<O(log(N))
Relevancy	Less. Since this algorithm compute rank at indexing time.	Less as ranking is based on calculating weight at indexing time.
Quality of result	Medium	Higher than PR
Query Dependency	Query independent	Query independent
Advantage	Rank is calculated on the basis of the importance of a page.	Main advantage of this algorithm is that it assigns larger rank value to more important pages.
Disadvantage	The main disadvantage is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing web site.	It is based only on the popularity of the web page.

*N is no. of web pages

VII.CONCLUSION & FUTURE WORK

The usual search engines usually result in a large number of pages in response to user's queries, while the user always wants to get the best in a short span of time so he/she does not bother to navigate through all the pages to get the required ones. The page ranking algorithms play a major role in making the user search navigation easier in the results of a search engine. The PageRank and Weighted Page Rank algorithm give importance to links rather than the content of the pages. According to PageRank algorithm, rank score of a web page is divided evenly over the pages to which it links whereas Weighted PageRank algorithm assigns larger rank values to more important (popular) pages. The PageRank and WPR return the important pages on the top of the result list. Some of the future work in these algorithms includes the following:

- They can be combined with some content based ranking algorithm to improve relevancy of the web pages.
- The concept of no. of visits of a link can also be used in calculating weight of page in case of weighted pageRank algorithm.

REFERENCES

1. N. Duhan, A. K. Sharma and Bhatia K. K.,' Page Ranking Algorithms: A Survey', Proceedings of the IEEE International Conference on Advance Computing, 2009, 978-1-4244-1888-6.

2. S. Brin, and Page L., 'The Anatomy of a Large Scale Hypertextual Web Search Engine', Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

3. Larry Page, and Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bring Order to the ' , Technical report in Stanford U, 1998.

4. Etizioni O.,'The World Wide Web: Quagmire or Gold Mine, Communications of the ACM, 39(11)', pp. 65-68(1996).

5. Wenpu Xing and Ghorbani Ali, 'Weighted PageRank Algorithm', Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

6. G.Kumar; N. Duhan; A.K. Sharma, 'Page Ranking Based on Number of Visits of Links of Web Page ', International Conference on Computer & Communication Technology (ICCCT), 2011.

7. R. Kosala.; H.Blockeel ,'Web Mining Research: A survey', In ACM SIGKDD Explorations, 2(1), PP. 1–15.

8. Mercy Paul Selvan ,A .Chandra Sekar and A.Priya Dharshin , 'Survey on Web Page Ranking Algorithms' ,International [Journal of Computer Applications (0975 – 8887) Volume 41– No.19, March 2012.

9. J. Kleinberg, 'Authoritative Sources in a Hyper-Linked Environment' , Journal of the ACM 46(5), pp. 604-632,1999.