ISSN 2047-3338

A Comparative Analysis of Page Ranking Algorithms

Dr. Paras Nath Gupta[1], Pawan Singh[2], Punit Kr Singh[3] and Amit Kumar[4]

***Abstract*-- With the growth of World Wide Web the information containment is also increased. The search engines are used to retrieve the information from WWW so it is the responsibility of search engine to provide the relevant information to the web user against the query submitted to it. This paper describes the role of hyperlink structure as a graph and its characteristics used in searching. This paper explores different page ranking algorithm, factors affecting the page ranking and different problems in page ranking algorithms.**

***Index Terms*– Web Structure, Page Ranking, Web Graph and Hyperlink Structure**

I. INTRODUCTION

THE web is a collection of document pages and hyperlinks that interconnect them. Therefore the web is a graph. It is a directed labelled graph whose nodes are the document and the edges are the hyperlink between them. The web is a huge structure growing rapidly. This network of information lacks organization and structure, and is only held together by the hyperlink. In order to make the navigation easier people use search engines and focus their search by querying, using specific keywords. In the beginning the amount of information content was not huge so search engines uses manually made list covering popular topics.

An index is maintained, containing list of all the words and mapping the pages containing these words. This index is used answer the users query in search engines. To get the better results ranking on the basis of importance and relevance taken under consideration. The method is generated using the link structure of the web and considering the incoming and outgoing links of the page. There are number of methods generated using different characteristics of web. Broader [1] describe the web graph similar to giant bow as shown in Fig. 1. The structure is named as Web Microscopic Structure. The central core – Strongly Connected Component, is the heart of the web where all the pages can reach one another along directed links. The second part IN contains pages which can reach the central core but cannot be reached by SCC. The third part OUT is having pages which can be reached from SCC by they cannot reach SCC pages. The last part Tendrils contains the page that cannot reach and cannot be reached from SCC.

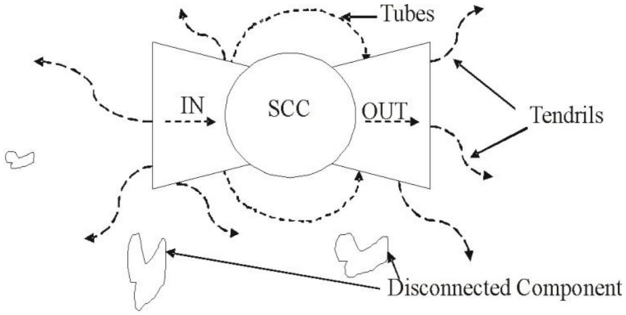


Fig. 1: Web Microscopic Structure

The web can be viewed as a directed labelled graph as shown in Fig. 2 whose nodes are the documents or pages ad the edges are hyperlinks between them. Let the graph Web Graph is G having two sets V and E.

V: set of vertices, finite and nonempty set
E: pair of vertices called edges

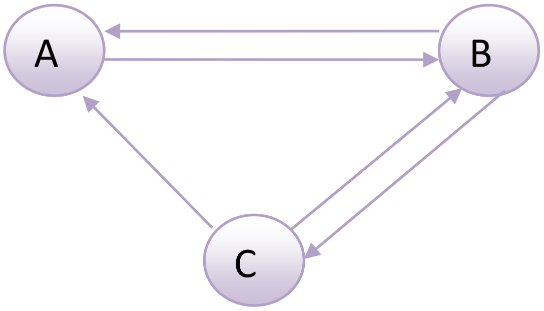


Fig. 2: A Directed Graph G

[1] Associate Professor, Department of Mathematics, A.M. College, Gaya (Email: drparasnathgupta.magadhuni@yahoo.com)

[2] Assistant Professor, Department of Computer Science, IIMT Engineering College, Meerut (Email: pawansingh3@yahoo.com)

[3] Assistant Professor, Department of Computer Science, IIMT Engineering College, Meerut (Email: punit20.singh@live.com)

[4] Assistant Professor, Department of Computer Science, IIMT IET, Meerut (Email: amit_m1@rediffmail.com)

V(G) = {A,B,C}
E(G) = {(A,B),(B,A),(B,C),(C,B),(C,A)}

II. WEB PAGE RANKING ALGORITHMS

Different search engines have developed the page ranking, taking under consideration the relevance, importance and content scores to mine the information and order them according to user interests. Some of the popular page ranking algorithms are discussed in the following section.

A. Page Rank Algorithms

S. Brin and Larry Page [2] developed a ranking formula utilized by Google, named Page Rank after Larry Page , that uses the link structure of the WWW to see the importance of pages of web sites. This formula states that if a page has some vital incoming links to that then its outgoing links to alternative pages additionally become vital. Therefore, it takes back links into consideration and propagates the ranking through links. Thus, a page incorporates a high rank if the addition of the ranks of its back links is high. The Page Rank formula assigns a Page Rank score to over twenty five billion pages [3] on the internet. A simplified version of Page Rank is outlined in eq. (1):

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (1)$$

Where u represents an internet web page, B(u) is that the set of pages that spot to u. PR(u) and PR(v) are the rank scores of page u and v, correspondingly. Nv denotes the quantity of outgoing links of page v, c may be a factor used for normalization. In Page Rank, the rank score of a page, p, is equally divided among its outgoing links. The values appointed to the outgoing links of page p are in tern used to successively calculate the ranks of the pages to that page p are pointing. AN example showing the distribution of page ranks is illustrated in Fig. 3.

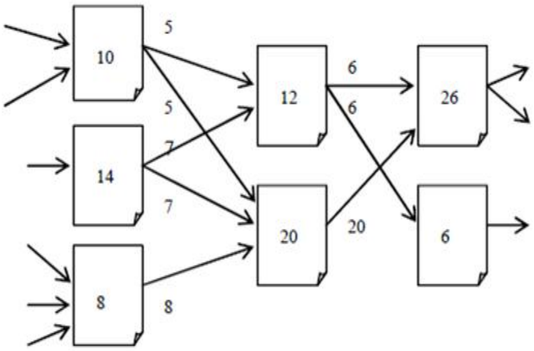


Fig. 3: Distribution of Page Rank

The modified version is given as:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (2)$$

where d is a dampening factor that is generally set to 0.85. d may be thought of as the probability of users' following the links and could regard (1 − d) as the page rank allocation from non-directly linked pages.To explain the functioning of Page Rank, let us take an example hyperlinked structure exposed in Figure 4, where A, B and C are three web pages. The Page Ranks for pages A, B, C are calculated by using eq. (2).

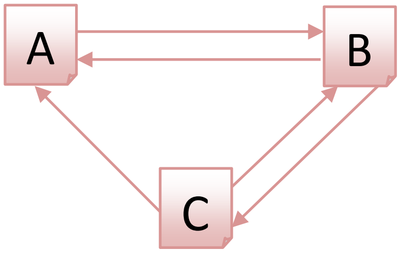


Fig. 4: Example Hyperlinked Structure

PR(A) = (1-d) + d((PR(B)/2+PR(C)/1) (2a)
PR(B) = (1-d) + d(PR(A)/2+PR(C)/1) (2b)
PR(C) = (1-d) + d(PR(B)/2) (2c)

By calculating the above equations with d = 0.5, the page ranks of pages A, B and C becomes:
PR(A)=1.2, PR(B)=1.2, PR(C)=0.8.

B. *Weighted Page Rank Algorithm*

Wenpu Xing and Ali Ghorbani [4] suggested an expansion to standard web Page Rank called Weighted Page Rank (WPR). This algorithm allocates larger rank values to more vital pages instead of distributing the rank value of a web page equally among its outgoing linked web pages. Every out-link web page gets a value according to its popularity. The popularity of a web page is computed from the number of in-links and out-links as Win (v,u) and Wout (v,u), accordingly. Win (v,u) (as in eq. (3)) is the weight of web link(v, u) computed on the basis of the count of in-links of web page u and the count of in-links of all reference web pages of web page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \qquad (3)$$

Iu and Ip represents the count of in-links of web page u and web page p, correspondingly. Wherever R(v) provides the reference web page list of web page v and Wout (v,u) (as in eq. (4)) represents the weight of link(v, u) computed based on the count of out-links of page u and the count of out-links of all reference web pages of page v.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \qquad (4)$$

where Ou and Op provides the count of out-links of web page u and web page p, accordingly. Considering the importance of web pages, the actual Page Rank formula eq. (2) is modified as given in eq. (5).

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \quad (5)$$

To illustrate the working of WPR refer again to Figure 4.then the PageRank equations become:

$$PR(A) = (1-d) + d\,(PR(B).W^{in}_{(B,A)}.W^{out}_{(B,A)} + PR(C).W^{in}_{(C,A)}.W^{out}_{(C,A)})$$

$$PR(B) = (1-d) + d\,(PR(A).W^{in}_{(A,B)}.W^{out}_{(A,B)} + PR(C).W^{in}_{(C,B)}.W^{out}_{(C,B)})$$

$$PR(C) = (1-d) + d\,(PR(B).W^{in}_{(B,C)}.W^{out}_{(B,C)})$$

The weights of in-links as well as out-links can be calculated as:

W^{in} (B,A) = IA/(IA+IC) = 2/(2+1) = 2/3
W^{out} (B,A) = OA/(OA+OC) = 1/(1+2) = 1/3

Similarly other values after calculation are:

W^{in} (C,A)=1/2 and W^{out} (C,A)=1/3
W^{in} (A,B)=1 and W^{out} (A,B)=1
W^{in} (C,B)=1/2 and W^{out} (C,B)=2/3
W^{in} (B,C)=1/3 and W^{out} (B,C)=2/3

After substituting d= 0.5 and above calculated weight values, (5a), (5b) and (5c) become:

PR(A) = 0.65 , PR(B) = 0.93, PR(C) = 0.60

Here $PR(B) > PR(A) > PR(C)$.

The resulting order of pages provided by Page Rank and WPR is different. To compare the WPR with original Page Rank, the authors categorized the resultant pages of a query into four classes based on their relevancy to the given query:

i). Very Relevant pages (VR
ii). Relevant pages (R)
iii). Weak-Relevant pages (WR)
iv). Irrelevant pages (IR)

Relevancy Rule: The relevancy of a page depends on its category and its position in the page-list. The larger the relevancy value is, the better is the result. The relevancy κ of a page-list is a function of its category and position:

$$\kappa = \sum_{i \in R(p)} (n-i) \times W_i \qquad (6)$$

where i represents the ith web page in the end result web page-list R(p) and n represents the first n web pages selected from theend result web page list and W_i gives the weight of page i. W_i = {v1, v2, v3, v4}, where v1, v2, v3 and v4 are the values assigned to a page if page is VR, R, WR and IR respectively. Also the values are chosen in such a way so that v1 > v2 >v3 >v4. The value of W_i for an experiment could be decided through experimental studies.

C. Page Ranking based on Links Visits (PRLV)

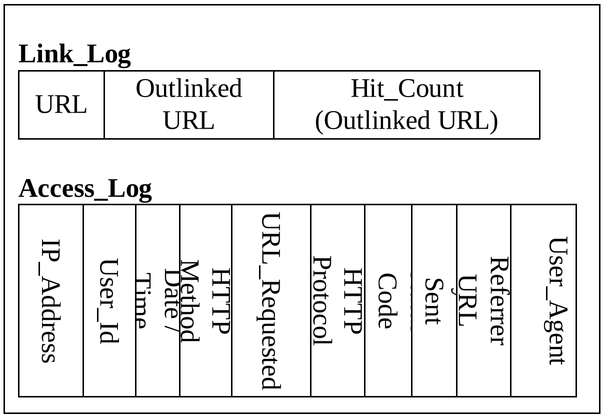
PRLV (Page Ranking based on Link Visits) [5] based on Web Structure Mining and Usage Mining is taking the user visits of pages/links to determine the importance and relevance score of the web pages. To accomplish the task many subtasks are performed as:

1). Storage of user's access information (hits) on an outgoing link of a page in related server log files.
2). Fetching of pages and their access information by the targeted web crawler.
3). For each page link, computation of weights based on the probabilities of their being visited by the users.
4). Final rank computation of pages based on the weights of their incoming links.
5). Retrieval of ranked pages corresponding to user queries.

The weights are determined for out-links of pages and ranks are computed by taking back-links into account.

Calculation of Visits (hits) of links: The hit count of each hyperlink is also stored, which is calculated easily by counting the distinct IP addresses visiting the corresponding page. The format of log files stored on the web server is shown in Figure 5. The link_log contains information about the page URLs, their hyperlinks and total hit count of each hyperlinked URL. The count of visits stored in link_log is computed from the second log called access_log , it has the format of the NCSA Combined Log Format [5]. By processing the access_log and counting the distinct IP addresses or User_Ids visiting the URL & stored in link_log.

Calculation of Rank Score: The search engines index most of the page information as periodically accessing from different web servers databases by the crawlers at the time of crawling. The working of crawlers for PRLV is to fetch the



Link_Log

URL	Outlinked URL	Hit_Count (Outlinked URL)

Access_Log

IP_Address	User_Id	Date / Time	HTTP Method	URL_Requested	HTTP Protocol	Code	Sent	Referrer URL	User_Agent

Fig. 5: Format of Server Log Files

pages as well as their hit counts stored in link_logs and sent to the PRLV Calculator. Every link in the crawled web graph is assigned a weight, indicating its probability of being visited by the users. Thus, ranking is propagated iteratively through back linked pages. Some terminologies used are:

Outbound link, a page p having n hyperlinks in it and the set of outbound link is denoted by: O(p)= { o_1, o_2,…o_n }, where each oi is a web page which can be accessed from web page p.

Inbound Link to page p is a set B of m inbound links if: B= {b_1, b_2,…b_m} where each b_i is a web page from which web page p can be accessed .

Probability-Weight of Link - For a web page p O(p) is a outbound hyper link set where every outbound hyper link is attached a integer value Visit Count (VC).Then the weight for link between p and o web pages is given by eq. (7).

$$Weight_{link}\ (p, o) = \frac{VC\,(p, o)}{\sum_{o\ \epsilon O(p)} VC\ (p, o)} \quad (7)$$

As per the example hyperlinked structure shown in Fig. 6, where the constant on each link indicates the visit count and a value in brackets indicates the calculated weight. The weight of link (D, F) is:

$$Weight_{link}\ (D, F) = \frac{100}{100 + 75 + 25} = \frac{1}{2}$$

Weight $_{link}$ (D, G) = 3 / 8 ,Weight $_{link}$ (D, H) = 1/ 8

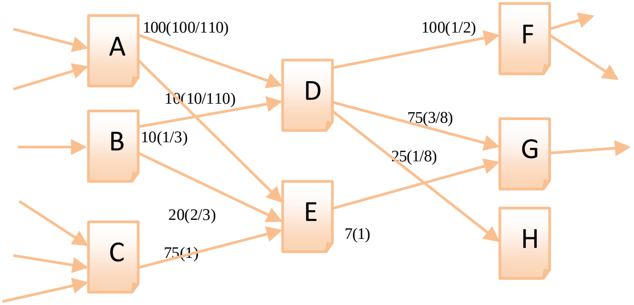


Fig. 6: Unequal Distribution of Link Weight

Page Rank based on Link Visits (PRLV), If p is a page having inbound-linked pages in set B(p), then the rank (PRLV) is given by eq. (8):

$$PRLV(p) = (1 - d) + d \sum_{b \in B(p)} PRLV(b).Weight_{link}\,(b, p) \quad (8)$$

where d is the damping factor as is used in PageRank, $Weight_{link}$ (.) is the weight of the link. The ranks for pages A, B and C are calculated. Following equations are obtained after substituting the values as indicated and calculated in above example.

PRLV(A)= (1-d) + d (PRLV(B).3/4+PRLV(C).1/3) (9)

PRLV(B) = (1-d) + d(PRLV(A).1+PRLV(C).2/3) (10)

PRLV(C) = (1-d) + d (PRLV(B).1/4) (11)

Taking d = 0.5, these equations can easily be solved using iteration method and the final results obtained are:

PRLV(A) = 1.08, PRLV(B) = 1.26, PRLV(C) = 0.66

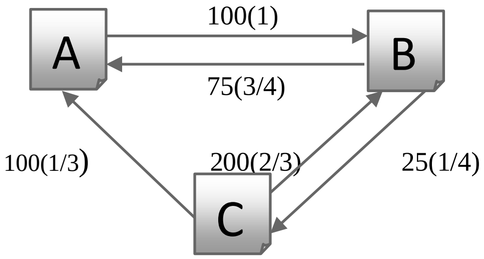


Fig. 7: Hyper linked structure

The PRLV calculates the ranks of the pages with the help of information collected from web crawler and passes these values to the web crawler.

D. *Web Page Ranking using Link Attributes*

Suppose that a user surfing the Web randomly and user is in a page, with certain probability. Ricardo [6] gives a variant WLRank assigns the ranking value R(i) to page I using the following eq. (12).

$$R(i) = \frac{q}{T} + (1 - q) \sum_{j} \frac{W(j, k)R(j)}{\sum_k W(j, k)} \quad (12)$$

where T is the total number of pages ,q is the probability (damping factor) of leaving page p and

W(j, i) = L(j, i) (c + T(j, i) + AL(j, i) + RP(j, i)) (13)

where given a link from page j to page i we have:

- L(j, i) is 1 if the link exists, or 0 otherwise, and c is a constant that gives a base weight to every link,
- T(j, i) is a value that depends on the tag where the link is inserted,
- AL(j, i) is the length of the anchor text of the link divided by a constant d that depends that estimates the average anchor text length in characters, and
- RP(j, i) is the relative position of the link in the page weighted by a constant b.

The term T(j, i) may be a sequence of constants depending on the tag wherever the link is, if the link is within a < h1> tag, a high T(j, i) price is appointed and slightly less for < h2>, etc.

AL (j, i) offers a lot of worth to links wherever the creator explained in additional detail what web resource is being linked. RP (j, i) offers a lot of weight to links that square measure at the start of the page rather that at the last of the page.

E. *Time Rank: Based on visiting time*

The Time Rank model [7] is generated by inducing time based visiting model. If the page content is related to keywords of user query then the user will stay on page for a long period, the other user will leave quickly giving short visiting time to the page. In Time Rank, each page has n-ranking scores which is based on n-number of topic, so are TSPR (i), which is calculated offline. Next the user's keywords submitted in query are matched against the number of topics. The relative probability between query q and topic i is suggested on the basis of Bayesian theorem [8] as in eq. (16):

$$P_r\ (T(i) \mid q) = \frac{\Pr(q, T(i))}{\Pr(q)} \quad (14)$$

$$= \frac{\Pr(T(i)) * \Pr\ (q \mid T(i))}{\Pr\ (q)} \quad (15)$$

$$\approx \Pr(T(i)) * \Pr(q \mid T(i)) \quad (16)$$

Ti → is the topic i of each page.
P_r (T(i)) → means the proportion of pages related to topic i in the whole pages set.
P_r (Ti | q) → means the probability of the query q belonging to topic i.

The topic sensitive page rank used in the Time Rank is given by:

$$TPSRt(j) = \alpha \sum_{i \in B} \frac{TSPR(i)}{|\ Fi\ |} + (1 - \alpha).Et(i) \quad (17)$$

Single jump probability 1/n is replaced byE_t= {E(1), E(2)............ E(n)}, n is the no. of topics.

$$E(i) = \begin{cases} 1/n_t \\ 0 \end{cases}$$

n_t → number of pages related to topic.

There are n TSPR scores corresponding to topics .It is calculated statically offline. After some running time of search engines the time vector related to topics for every page can be calculated and hence every page is assigned as page rank based on time visited as in eq. (18).

$$TIMEPRt(j) = TSPRt(j) * T(t) \quad (18)$$

Where time vector Tv = {T(1), T(2), ...T(n)}

T(i) - user's total visiting time of a page related to topic i. Time rank means that irrespective of similarity of similar link structure of two web pages, the page having longer visited time gets the high score.

III. OUTCOME OF COMPARISON OF VARIOUS WEB PAGE RANKING BASED ALGORTHMS

Based on the literature analysis ,some of the page ranking based algorithms are compared in Table 1 to compare them some parameters are selected as methodology, I/p parameters, Output Result , relevancy and limitations.

A. *Factors Affecting the Ranking*

The finding of the survey is that the generated different page ranking algorithms works on some specific characteristics or factors of the web structure (i.e. graph) and the behaviour of the user with reference to surfing the internet. There are different finite factors being used by the page ranking algorithms in crawler to help the users finding the relevant pages as per their requirement which is submitted to search engine in the form of query. These factors are as follows:

a) Incoming links/outgoing links: The links from and to the pages is taken under consideration as link coming from important page suggest the important page and outgoing links takes the weightage to other pages. Thus the incoming & outgoing links effects the importance of pages.

b) Content: Obviously the user approach is to find the relevant page containing the required content. So the content on a page itself reflects its importance and it is notified by the words used in page content as well as there frequency.

c) Popularity: For a topic there may be number of pages related to it. The page visited by maximum number of users (most popular among users) is considered the most relevant.

d) Page Generation Time: One should take care about the time of page generation as the older page with less relevancy may get more weightage (Rank) than new one with better relevancy.

Table 1: Comparison of Page Rank based algorithms

	Page Rank	**Weighted Page Rank**	**PRLV**	**Web Page Ranking using Link Attributes**	**Time Rank**
Technique Used	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Usage Mining	Web Structure Mining, Web Content Mining	Web Usages Mining
Methodology	This algorithm calculates the score for pages during indexing of the pages	Weight of web page is calculated on the premise of input and outgoing links and on the premise of weight the importance of page is determined.	Computes scores at indexing time. Pages are sorted according to importance and connexion.	It offers completely different weight to web links supported by 3 attributes: Relative position in page, tag wherever link is contained, length of anchor text.	In this algorithm the visiting time is appended to the calculated score of the actual page rank of that page.
I/p Parameters	Back links	Back links and Forward links.	Inbound link, outbound link and visit counts of links	Content, Back and Forward links	Original Page Rank and Sever Log
O/p Quality	Medium	High	High	Medium	Medium
Relevancy Level	Less	Less	Moderate	Moderate	High
Importance	High	High	High	Not specifically quoted	High
Nature of Rank	Less dynamic (rank changes with link structure)	Less dynamic	More dynamic (rank changes with visit counts & structure of links)	Less Dynamic	More dynamic (rank changes with duration of use & structure of links)
Limitation	Results come back at the time of indexing and not at the query time.	Relevancy is ignored.	Extra effort on crawlers to fetch the visit counts of pages from web servers. Extra calculations to seek out the weights of links.	Relative position wasn't so effective, indicating that the logical position not continuously matches the physical position.	Important pages are ignored as a result of it increases the rank of those sites that are opened for long time.

e) Duration of page used: The calculation of relevancy and rank must include not only the number of times the page is used but also the duration of page used. A user will stay on a page for a long time only if it is as per his desire otherwise he or she may leave the page soon.

f) User Type: Different users have different approach, view and prospects. By categorizing the users it may be helpful to find the required page with content.

g) Relative position of a tag for a hyperlink in the page: The page may have number of links located at different places on the sane page in the form of tag. A tag at the beginning of a tag is given more weight age with respect to the tags at the end.

h) Length of the anchor text: In this case the more weight is given to the links where the creator explained in great detail what web resource is being linked. i) Number of link traversed: If a user has to traverse a long distance in the web (gone through more number of links) than the destination page is having low rank i.e. the rank is inversely proportional to the number of links required to traverse to reach the desired page in the web graph.

B. Problems in Page Ranking Algorithms

a) Rich-get richer: The popular high rank web pages become more and more popular and the young high quality pages are not picked by the ranking algorithms.

b) Rank sink: A web owner creates a large number of bogus web pages all pointing to and pointed by a single target page. In this case the page rank algorithm assigns a higher ranking score to target than it deserves.

c) 0-1 gap problem: Many web pages do not have any out links and many web applications only consider a sub graph of the whole web. Even if a page has out links it might have been removed when the whole web is projected to a sub graph. Removing all the pages without out links is not a problem because it generates new zero-out-link pages. The probability of jumping to random page is 1 in zero-out-link page, but it drops to λ (λ=.15) for a page with single out- link there is a big difference between zero &one link out.

d) Topic drift: It means that while searching a page asked by the user query in well connected pages a ranking algorithm makes a traversal towards the non very relevant pages.

e) In some algorithms if a new page is inserted or added than it takes large calculations.

III. CONCLUSION AND FUTURE SCOPE

This paper describes the working of different page ranking algorithms used to retrieve the relevant pages through search engines. These algorithms are been compared on the basis of literature survey and the centre of attraction the important factors affecting the ranking of the web pages are suggested as well as the attention is drawn towards the problem arisen in the working of different page ranking algorithms. These factors can be considered to generate a new algorithm in future to eliminate the problems.

REFERENCES

[1] A. Broder, R. Kumar, F Maghoul, P. Raghavan, S. Rajagopalan, R.Stata, A. Tomkins, J. Wiener, “Graph Structure in the Web”, Computer Networks: The International Journal of Computer and telecommunications Networking, Vol. 33, Issue 1-6, pp 309-320, 2000.

[2] S. Brin and L. Page (1998). The Anatomy of a large-scale hypertextual Web search engine. In seventh international World Wide Web Conference, Brishbane, Australia, 1998.

[3] Amy N. Langville and Carl D. Meyer, Deeper Inside PageRank, October 20, 2004.

[4] Wenpu Xing, Ali Ghorbani. Weighted PageRank Algorithm[C], Proceedings of the Second Annual Conference on Communication Networks and Services R -esearch, IEEE, 2004.

[5] Sharma, A.K., Duhan, N. and Kumar, G. (2010) A Novel Page Ranking Method based on Link- Visits of Web Pages. International Journal of Recent Trends in Engineering and Technology, Vol. 4, No. 1, pp 58-63.

[6] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329,2004

[7] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[8] Dimitri, P. Betsekas and John N. Tsitsiklis, Introduction to Probability. Athena Scientific, 2002.