Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page

Neelam Tyagi, Simple Sharma

Abstract: ***The World Wide Web consists billions of web pages and hugs amount of information available within web pages. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture. When a user refers a query to the search engine, it generally returns a large number of pages in response to user's query. To support the users to navigate in the result list, various ranking methods are applied on the search results. Most of the ranking algorithms which are given in the literature are either link or content oriented. Which do not consider user usage trends. In this paper, a page ranking mechanism called Weighted PageRank Algorithm based on Visits of Links (VOL) is being devised for search engines, which works on the basis of weighted pagerank algorithm and takes number of visits of inbound links of web pages into account. The original Weighted PageRank algorithm (WPR) is an extension to the standard PageRank algorithm. WPR takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. The proposed algorithm is used to find more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale.***
The paper also presents the comparison between original and VOL method.

Keywords: World Wide Web, Search Engine, PageRank, Inbound Link, Outbound Link.

I. INTRODUCTION

The World Wide Web (Web) is popular and interactive medium to propagate information today. The Web is huge, diverse, dynamic, widely distributed global information service center. As on today WWW is the largest information repository for knowledge reference.

With the rapid growth of the Web, users get easily lost in the rich hyperlink 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. When a user makes a query from searchengine, it generally returns a large number of pages in response to user queries. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of relevant pages in the search result-list. To assist the users to navigate in the result list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. In that way user can find the most important and useful result first. There are a variety of algorithms developed, few of them are PageRank, HITS, SALSA, RANDOMZE HITS, SUBSPACE HITS, SIMRANK etc. As most of the ranking algorithms proposed are either link or content oriented in which consideration of user usage trends are not available. In this paper, a page ranking mechanism called Weighted PageRank Algorithm based on Visits of Links (VOL) is being devised for search engines, which works on the basis of Weighted PageRank Algorithm and takes number of visits of inbound links of web pages into account. The original Weighted PageRank algorithm (WPR) is an extension to the standard PageRank algorithm. WPR takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages.

The main purpose of the proposed algorithm is finding more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale. The details of the proposed algorithm will clarify later in this paper. The rest of this paper is organized as follows: a brief summary of related work is given in Section II. Section III describes the proposed algorithm in detail. The result analysis is given in Section IV. Section V summarizes the results and draws a general conclusion.

II. RELATED WORK

Brin and Page [2] developed PageRank algorithm at Stanford University based on the hyper link structure. PageRank algorithm is used by the famous search engine, Google. PageRank algorithm is the most frequently used algorithm for ranking billions of web pages. During the processing of a query, Google's search algorithm combines precomputed PageRank scores with text matching scores to obtain an overall ranking score for each web page. Functioning of the Page Rank algorithm depends upon link structure of the web pages. The PageRank algorithm is based on the concepts that if a page surrounds important links towards it then the links of this page near the other page are also to be believed as imperative pages. The Page Rank imitate on the back link in deciding the rank score. Thus, a page gets hold of a high rank if the addition of the ranks of its back links is high. A simplified version of PageRank is given in Eq. 1:

Manuscript Received on June 03, 2012.

Neelam Tyagi, Computer Science & Engineering, Manav Rachna International University, Ballabgarh, India

Simple Sharma, Asst. Prof.,Computer Science & Engineering, Manav Rachna International University, Faridabad, India

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

Where u represents a web page, B (u) is the set of pages that point to u, PR (u) and PR (v) are rank scores of page u and v respectively, N_v indicates the number of outgoing links of page v, c is a factor applied for normalization.

Later PageRank was customized observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2:

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

Where d is a dampening factor that is frequently set to 0.85. d can be thought of as the prospect of users' following the direct links and (1 –d) as the page rank distribution from non- directly linked pages.

Wenpu Xing et. al.[8] discussed a new approach known as weighted pagerank algorithm (WPR). This algorithm is an extension of PageRank algorithm. 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. WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query.

According to author the more popular webpages are the more linkages that other webpages tend to have to them or are linked to by them. The proposed extended PageRank algorithm–a Weighted PageRank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$, respectively.

$W^{in}_{(v,u)}$ given in eq. (3) is the weight of *link*(*v, u*) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*.

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

Where I_u and I_p represent the number of inlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*. $W^{out}_{(v,u)}$ given in eq. (4) is the weight of *link*(*v, u*) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v*.

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

Where O_u and O_p represent the number of outlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

Considering the importance of pages, the original PageRank formula is modified in eq. (5) as

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

Gyanendra Kumar et. al. [6] proposed a new algorithm in which they considered user's browsing behaviour. As most of the ranking algorithms proposed are either link or content oriented in which consideration of user usage trends are not available. In this paper, a page ranking mechanism called Page Ranking based on Visits of Links(VOL) is being devised for search engines, which works on the basic ranking algorithm of Google, i.e. PageRank and takes number of visits of inbound links of web pages into account. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale.

In this paper as the author describe that in the original PageRank algorithm, the rank score of page p, is evenly divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page, p. So, he proposed an improved PageRank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculated based on visits of inbound links.

The modified version based on VOL is given in equation (6)

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} \frac{L_u \, PR(v)}{TL(v)} \qquad (6)$$

Notations are :

- d is a dampening factor ,
- u represents a web page,
- B(u) is the set of pages that point to u,
- PR(u) and PR(v) are rank scores of page u and v respectively,
- L_u is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v.

III. WEIGHTED PAGERANK BASED ON VISITS OF LINKS(VOL)

We have seen that the original Weighted PageRank algorithm assigns larger rank values to more important (popular) pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$, respectively. Here we proposed an improved Weighted PageRank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks. We do not consider here the popularity of outlinks which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links.

The modified version based on WPR (VOL) is given in eq. (7)

$$WPR_{vol}(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{L_u \, WPR_{vol}(v) W^{in}_{(v,u)}}{TL(v)} \quad (7)$$

Notations are :

- d is a dampening factor ,
- u represents a web page,
- B(u) is the set of pages that point to u,
- WPR_{VOL} (u) and WPR_{VOL} (v) are rank scores of page u and v respectively,
- L_u is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v.

A. Algorithm

The various steps of the proposed algorithm are given below**:**

Step 1: Finding a Website: Find a website which have rich hyperlinks because the weighted PageRank and WPR (VOL) methods rely on the web structures.

Step 2: Building a Web Map: Then generate the web map from the selected website.

Step 3: Calculate W^{in} (v,u): Then calculate the W^{in} (v,u) for each node present in web graph by applying the equation (3) as below.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \epsilon R(m)} I_p} \quad (3)$$

Where

- $W^{in}_{(v,u)}$ is the weight of *link*(*v, u*) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*.
- I_n and I_p are the number of incoming links of page n and page p respectively.
- R (m) denotes the reference page list of page m.

Apply proposed formula: Now calculate the PageRank value of the nodes present in web graph by using the proposed formula (8)

$$WPR_{vol}(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{L_u \, WPR_{vol}(v) W^{in}_{(v,u)}}{TL(v)} \quad (8)$$

Where

- u represents a web page,
- B(u) is the set of pages that point to u,
- d, is the dampening factor.
- WPR_{vol} (u) and WPR_{vol} (v) are rank scores of page u and v respectively,
- L_u denotes number of visits of link which is pointing page u form v.
- TL(v) denotes total number of visits of all links present on v.

Step 5: Repeat by going to step 4: final step will be used recursively until the values are to be stable.

B. Example Illustrating Working of Weighted PageRank Based on VOL

To explain the working of original and proposed Weighted PageRank algorithm, let us take an example hyperlinked structure shown in fig 1, we regard a small web graph consisting of three pages A, B and C. where page A links to the page B and C, page B links to page C and page C links to page A and each link has its corresponding visits.

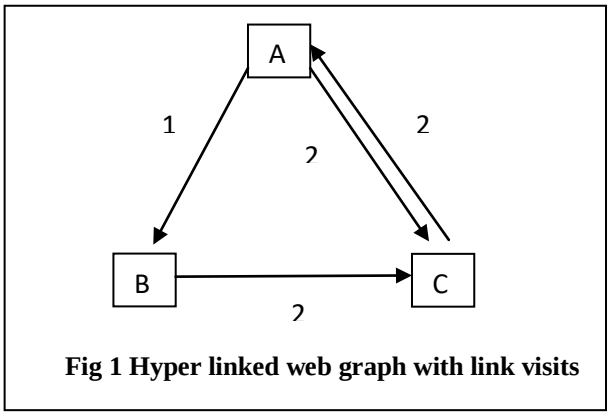

Fig 1 Hyper linked web graph with link visits

C. Calculation of Proposed Weighted PageRank Using VOL

The Proposed Weighted PageRank for pages A, B and C are calculated by using equation (1) as given above:

WPR_{VOL} (A) = (1-d) + d (WPR_{VOL} (C) $W^{in}_{(C,A)}$ L_A / TL(C))

WPR_{VOL} (B) = (1-d) + d (WPR_{VOL} (A) $W^{in}_{(A,B)}$ L_B / TL (A))

WPR_{VOL} (C)=(1-d) + d (WPR_{VOL} (A)$W^{in}_{(A,C)}L_C$ / TL(A)+ WPR_{VOL} (B)$W^{in}_{(B,C)}$ L_C / TL(B))

$$W^{in}_{(C,A)} = \frac{I_A}{I_A + I_B} = \frac{1}{1+1} = \frac{1}{2}$$

$$\frac{L_A}{TL(C)} = \frac{2}{2} = 1$$

$$W^{in}_{(A,B)} = \frac{I_B}{I_C} = \frac{1}{2}$$

$$\frac{L_B}{TL(A)} = \frac{1}{3}$$

$$W^{in}_{(A,C)} = \frac{I_C}{I_C} = \frac{2}{2} = 1$$

$$\frac{L_C}{TL(A)} = \frac{2}{3}$$

$$W^{in}_{(B,C)} = \frac{I_C}{I_A} = \frac{2}{1} = 2$$

$$\frac{L_C}{TL(B)} = \frac{2}{2} = 1$$

Put all these values in all above equations and calculate proposed WPR(VOL) value for each web page A, B, C.

I. Consider d=0.35

WPR_{VOL}(A) = 0.65+0.35 (1* $\frac{1}{2}$ * **1**) = 0.825

WPR_{VOL} (B) = 0.65+0.35 (0.825 * $\frac{1}{2}$ * $\frac{1}{3}$) = 0.698125

WPR_{VOL}(C) = 0.65+0.35 (0.825 * 1 * $\frac{2}{3}$ +0.698125*2*1) = 1.3311875

Calculate all these values iteratively until values become stable. Then we get the following iterative table 1:

Table 1 shows WPR_{VOL} values at d=0.35

WPR_{VOL} (A)	WPR_{VOL} (B)	WPR_{VOL} (C)
0.825	0.698125	1.3311875
0.882957812	0.701505872	1.347077599

0.885738579	0.701668083	1.347839993
0.885871998	0.701675866	1.347876572
0.8858784	0.70167624	1.347878328

The rank score of web pages A, B, C are when **d=0.35: WPR_{VOL} (A)=0.885, WPR_{VOL} (B)=0.701 and WPR_{VOL} (C)= 1.347**

II. *Consider d=0.5*

WPR_{VOL} (A) = 0.5+0.5 $(1*\frac{1}{2}*\mathbf{1}) = 0.75$

WPR_{VOL}(B) = 0.5+0.5 $(0.75*\frac{1}{2}*\frac{1}{3}) = 0.5625$

WPR_{VOL} (C) = 0.5+0.5 $(0.75*1*\frac{2}{3}$ +0.5625*2*1) = 1.3125

Calculate all these values iteratively until values become stable. Then we get the following iterative table 2:

Table 2 shows WPR_{VOL} values at d=0.50

WPR_{VOL} (A)	WPR_{VOL} (B)	WPR_{VOL} (C)
0.75	0.5625	1.3125
0.828125	0.5690104	1.345052082
0.83626302	0.569688585	1.348442925
0.837110731	0.569759227	1.348796137
0.837199034	0.569766586	1.34883293

The rank score of web pages A, B, C are when **d=0.5: WPR_{VOL} (A)=0.837, WPR_{VOL} (B)=0.569 and WPR_{VOL} (C)= 1.348**

III. *Consider d=0.85*

WPR_{VOL} (A) = 0.15+0.85 $(1*\frac{1}{2}*\mathbf{1}) = 0.575$

WPR_{VOL} (B) = 0.15+0.85 $(0.575*\frac{1}{2}*\frac{1}{3}) = 0.231458333$

WPR_{VOL} (C) = 0.15+0.85 $(0.575*1*\frac{2}{3}$ +0.231458333*2*1) = 0.869312499

Calculate all these values iteratively until values become stable. Then we get the following iterative table 3:

Table 3 shows WPR_{VOL} values at d=0.85

d	**0.35**		**0.5**		**0.85**	
	WPR	WPR_{VOL}	WPR	WPR_{VOL}	WPR	WPR_{VOL}
A	0.879	0.885	0.83	0.837	0.48	0.491
B	0.726	0.701	0.6	0.569	0.25	0.219
C	1.31	1.347	1.31	1.348	0.79	0.801

The rank score of web pages A, B, C are when **d=0.85: WPR_{VOL} (A) =0.491, WPR_{VOL} (B) =0.219 and WPR_{VOL} (C) = 0.801**

IV. HOW TO CALCULATE HITS (VISITS) OF LINKS

Here, to count the hits or visits of an outgoing links on a web page a client side script is used. Whenever a web page is accessed the script will be loaded on the client side from web server. Script will monitor the click as well as keyboard event to occur. When a event occur and if that event will happen over hyperlink then it will send a message to web server with information of current web page and hyperlink. On server side a data base of log file will be used to record the web page id, hyperlinks of that page and hit count of hyperlinks. Hit count will incremented every time a hit occur on hyperlink.

The database or log files will accessed by crawler at the time of crawling. This (hit count) crawled information will be stored in search engine's database which is used to calculate the rank value of different web pages or documents.

V. RESULT ANALYSIS

In this section we will describe and explain some results we got from the work we have been doing. Here we have taken a hyperlinked web graph, shown in Fig 1 and calculated PageRank value of each page based on original weighted page rank algorithm and proposed algorithm i.e. based on number of visits of links (VOL).

A. Observation

Here, we calculate the page rank score of both the algorithms at various values of "d". The various values of "d" on which we calculate the pagerank score of both original and proposed algorithm are 0.35, 0.50, and 0.85. After calculation we have found a different rank score of three web pages A, B and C. All these are highly dependent on this dampening factor (d). Here we have a table 4 which shows rank score calculated by proposed algorithm at various values of "d".

Table 4 shows rank score of VOL calculated at various values of "d"

WPR_{VOL} (A)	WPR_{VOL} (B)	WPR_{VOL} (C)
0.575	0.231458	0.869312
0.519458	0.22359	0.824462
0.500396	0.220889	0.80907
0.493855	0.219963	0.803788
0.49161	0.219645	0.801975

By the use of table 4, we can give a graphical representation of proposed algorithm in fig 2. In this we can observe that the smaller the dampening factor's (d) value larger will be the pagerank score. And similarly larger the d's value gives smaller page rank score.

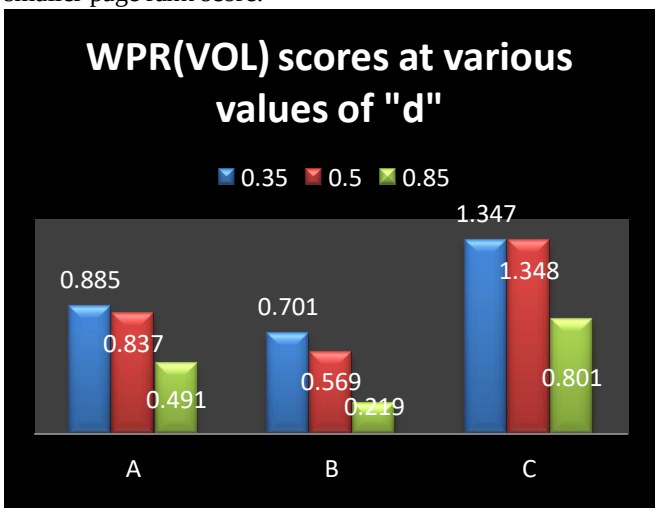

Fig 2 shows WPR (VOL) scores at various values of "d".

B. Observation

In this section we will show the variation between two algorithms i.e. original weighted pagerank algorithm and proposed weighted pagerank algorithm based on visits of links. Here, we will find out that the proposed algorithm is better than the original algorithm as the new algorithm will calculate more relevant web pages than that of existing one. As in the proposed algorithm we will assign more weightage to those web pages which is most visited by users, also provide more weightage to those web page which have higher popular inlinks. These two considerations will make the proposed algorithm performs better than original one. Now we have a table 8 which will show the comparison of two algorithms on the basis of rank scores at various values of "d".

Table 5 shows the rank score of WPR and WPR (VOL) at various values of "d".

d	A	B	C
0.35	0.885	0.701	1.347
0.5	0.837	0.569	1.348
0.85	0.491	0.219	0.801

By the use of this table 5, we can give a graphical representation in fig 23 which shows the variation between WPR (VOL) and WPR scores of different web pages of a web graph. In this we can observe that WPR (VOL) is performing better than original WPR. As the proposed algorithm calculates higher relevant score than the existing one.

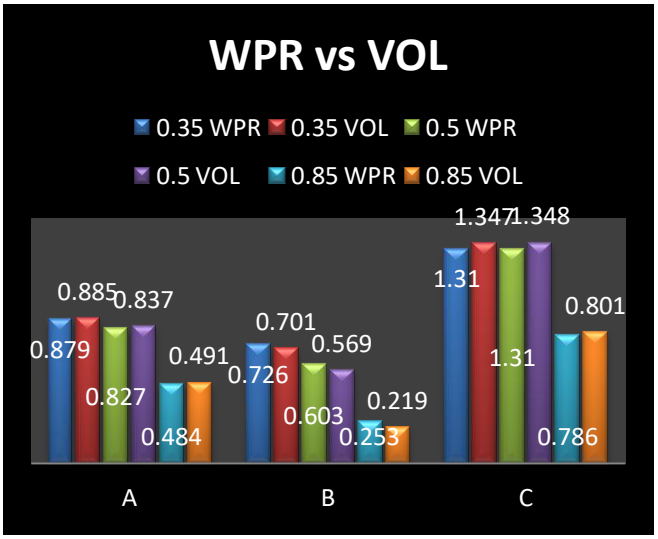

Fig 3 shows variation between VOL and WPR.

From this observation it has been proved that WPR (VOL) is far dynamic than original Weighted PageRank algorithm. It is also observed that the page which has more visits of incoming links is carrying more rank value than less visited pages. While original Weighted PageRank score of pages are depend upon the popularity of inbound and outbound links it does not consider visits of users.

VI. BENEFITS OF WEIGHTED PAGERANK BASED ON WPR (VOL)

WPR (VOL) supplies the most important web pages or information in front of users. Following are the advantages of WPR (VOL):

I. As WPR (VOL) method uses link structure of pages, the popularity of inlinks and their browsing information, the top returned pages in the result list is supposed to be highly relevant to the user information needs. A link with high probability of visit contributes more towards the rank of its out linked pages.

II. The rank value of any page by original Weighted PageRank method will be same either it is seen by user or not, because it is totally dependent upon link structure of Web graph and popularity of inlinks and outinks. While the ordering of pages using WPR (VOL) is more target-oriented.

III. In WPR(VOL), a user can not intentionally increase the rank of a page by visiting the page multiple times because the rank of the page depends on the probability of visits (not on the count of visits) on back linked pages.

The main issue to address is the periodic crawling of web servers so as to collect the accurate and up to date visit count of pages. Specialized crawlers need to be designed for fetching the required information of pages.

VII. CONCLUSION AND FUTURE SCOPE

In the above section a modified weighted page ranking algorithm is discussed which is more target-oriented than original weighted pagerank. This modified algorithm calculates PageRank value or importance of web pages based on the visits of incoming links on a page as well as the popularity of inlinks of a web page. It is not only consider link structure it includes the users focus on a particular page, but the main problem in this concept is calculation of visits of a links for that we have given a simple concept to monitor and count the hits or visits. User generally spends a lot of time in surfing through the search results to find the relevant pages. This modified algorithm provides more relevant results than original WPR. The ordering of pages in this way increases the relevancy of pages and thereof provides the user with quality search results.

Some of the futurework in this algorithmincludes the following:

a) The implementation can be done for the proposed concept to check the performance more accurately.

b) Currently proposed algorithm uses only link visit information from a user. One could think in other information like for example some feedback from search engine about which pages does the user choose from the whole list of results.

c) More experiments can be done with bigger set of data in order to be able to prove that the proposed algorithm is really more convenient that the existing ones.

d) Some improvements can be done in proposed method by adding some other experiments.

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 Web", Technical report in Stanford U, 1998.
4. R. Cooley, B. Mobasher, and Srivastava, J., "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 1997.
5. J. Kleinberg, "Hubs, Authorities and Communities", ACM Computing Surveys, 31(4), 1999.
6. Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.
7. R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
8. Wenpu Xing and Ghorbani Ali, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.