Web Search Engine
An integrated approach to rank webpages

Nikhil Kumar Singh, Sunny Sharan

Abstract— An average user spends most of his time to search some relevant document on the internet. The complexity of internet has increased by a very great margin. Existing web searching mechanism helps him to a certain extent, but they also throw some meaningless hundreds or thousands of webpages of which the user doesn't have any need or relevance. He still have to search and sort for the relative webpages that he actually require. Thus we require some extra information & mechanism to help users in the searching activity. We need to check the browsing habits and the number of visits to a particular website to determine the credibility of the search result return by the search engine. This is done without any tempering or hampering the current working of the user. Thus the information generated from this type of search engine is of high relevance and credibility.

Index Terms—Search Engine, Information Retrieval, Web Crawler Indexing, Page Ranking..

1 INTRODUCTION

In the eye-blink that has elapsed since the turn of the millennium the lives of those of us who work with information have been utterly transformed. Much most perhaps even all of what we need to know is on the World-Wide Web; if not today, then tomorrow. Intelligent Web Search Engine can be classified in three broad terms: - "intelligent, web, engine". An engine is anything that can be viewed as a perceiving the environment through sensors and returning or affecting the environment or taking any action on the environment through its effectors. For a web engine the environment is the web. It perceives it using words of an HTML document acquired from software sensors that connect to the internet using HTTP. The engine's actions will depend on the goal of the agent like for a search agent it would be seeking a website containing related information about the search string Web is a collection of thousands or millions of web pages. The web is where society keeps the sum total of human knowledge. It's where we learn and play, shop and do business, keep up with old friends and etc. A web engine can be said intelligent if it takes rational decision if it is presented with a problem. That means that if the engine is allotted a problem and a goal, then it will take actions which would lead to the successful completion of the goal.

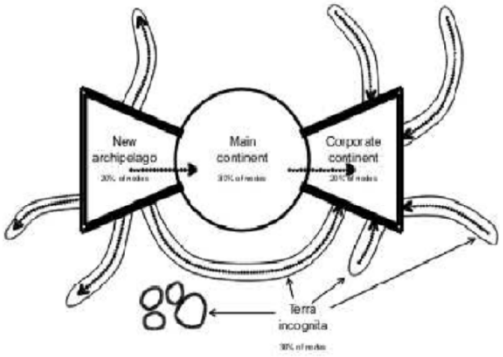
2 THE PROBLEM

The main problem that we are going to discuss in the research paper is how to provide the relevant information according to query of the user. The problem has gained a great amount of significance in this matter. As we know that information is just as equivalent as money and good information can help the user to achieve the goal in a very efficient manner. Every document present on the internet has an address in the form of a unique name or URL (Uniform Resource Locator). The problem is finding the document containing information one seeks is in finding the unique URL of it which of course is not easy to guess. Search engines like Google, Yahoo, MSN, etc helps the user in finding documents by fielding a query consisting of a few words representing the information which he wants to retrieve from the internet. The search engine matches the query to its own database of information about the web and returns documents containing words given in the query.

Now as the internet has grown too huge so the result returned normally consists of thousands of matching documents and it's impossible for the user to navigate through each of themTherefore in most cases nowadays, these search agents are unable to help the user by providing them documents matching their interests.

3 ARCHITECTURE OF WEB

The figure [7] shows the chart of the web, produced during the schematic study of its hyperlink structure. The shape is reminiscent of a bowtie.



The central knot is a giant subnet that we call the main continent, a large strongly connected structure in which all pages are linked together. Here you can travel from one page to any other by following the links {just as when surfing with a browser. This is where most surfing takes place.It's the dominant part of the web. The new archipelago is a large group of fragmented islands; the main continent can be reached from each island by following links. Most pages here are likely to be quite new, and haven't received many links.The corporate continent comprises pages that are all reachable from the main continent. Some are sinks, pages without any outward links at all (for example, most Word and PDF files); however, these are a minority. Though relatively

small compared to the main continent, these islands are significantly larger than those of the new continent. Terra incognita is the rest of the universe. The distinctive feature of its pages is that surfers who reach them haven't come from the main continent and won't get there in the future {though they may have travelled from the new archipelago, and they may end up in the corporate continent. Linked trails in terra incognita do not cross the main part of the web: the pages are simply disconnected from it.

4 Search Engine Architecture

The components of a search engine are: Web crawling (gathering webpages), indexing (representing and storing the information), retrieval (being able to retrieve documents relevant to user queries), and ranking the results in their order of relevance. Figure 1 presents a simplified view of the components of a search engine.

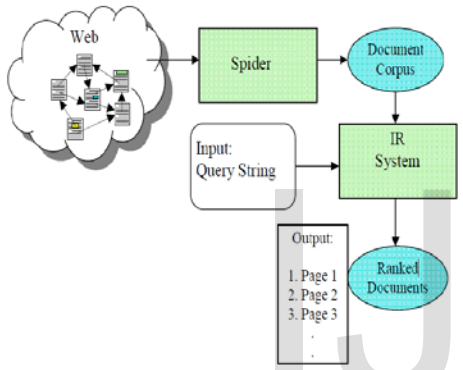


Figure 1. The simplified architecture of a search engine.

5 Present Ranking Algorithms

Brin and Page [3] 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 following equation:

$$PR(u) = c \times \sum_{v \in B(u)} (PR(v)/C(v))$$

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, C(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:

$$PR(u) = (1-d) + d \times \sum_{v \in B(u)} (PR(v)/C(v))$$

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 [1] 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 in(vu) and out(vu) , respectively. In(vu) given in following equation 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.

$$in(vu) = i(u) / \sum_{p \in R(v)} i(p)$$

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. out(vu) given in eq. Following equation 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.

$$outvu) = \frac{o(u)}{\sum_{p \in R(v)} o(p)}$$

Where Ou and Op 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:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} (PR(v) * in(vu) * out(vu))$$

Gyanendra Kumar [2] 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

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \left(\frac{l(u) * PR(v)}{tl(v)} \right)$$

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.

6 PROPOSED ALGORITHM

From the above algorithms , we observe that none of the above algorithm is perfect in every aspect,each uses its own parameters and criterias in order to rank the documents in order of their relevance . In case of pagerank algorithm , the criteria it uses in order to rank web pages is somewhat static in nature i.e it only considers the importance of a page based only on the pages that point to it(in links) . However , there must be a sense of dynamism in it [4] i.e. it must involve some amount of feedback from user(visits of webpages). So, our proposed algorithm comes into picture , In our proposed algorithm , we use two parameters in order to decide the relevance . The first criteria is the popularity of the page based on its outlinks. The second criteria is the number of visits by the user on the webpage. Here is what our proposed algorithm looks like. For a page "u" the weighted pagerank algorithm is given by:

$$wpr(u) = (1-d) + d \sum_{v \in B(u)} \left(\frac{l(uv)}{tl(v)} * out(vu) * wpr(v) \right)$$

Notations are :
d is a dampening factor ,
u represents a web page,
B(u) is the set of pages that point to u,
wpr(u) and wpr(v) are rank scores of page u and v respectively,
l(uv) 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.

In above algorithm we used the parameter out(uv) in order to measure popularity of a webpages based on the number of its outlinks, This out(uv) can be define as :

$$out(vu) = \frac{o(u)}{\sum_{p \in R(v)} o(p)}$$

Where
o(u) is number of outlinks of page(u)
$\sum_{p \in R(v)} o(p)$ is sum of outlinks of all pages that point to v

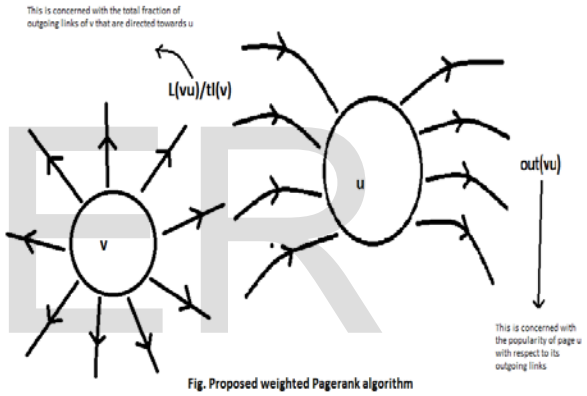


Fig. Proposed weighted Pagerank algorithm

7 IMPLEMENTATION

PAGERANK ALGORITHM :

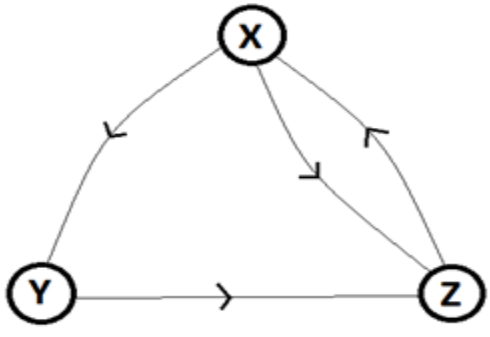


Fig: Pagerank Algorithm

Now, we discuss the implmentation of pagerank algorithm with respect to the above figure

We know ,the page rank of a page "u" is calculated using the following formulae:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \left(\frac{PR(v)}{C(v)}\right)$$

So, pagerank of page X is given as

PR(X)= (1-d)+ d (PR(Z)/C(Z))

Pagerank of page Y is given as

PR(Y)= (1-d)+ d (PR(X)/C(X))

Pagerank of page Z is given as

PR(u)= (1-d)+ d (PR(X)/C(X)+PR(Y)/C(Y))

From the above diagram we have

C(x)=2 // two outgoing edges from x

C(y)=1 //one outgoing edge from y

C(z)=1 //one outgoing edge from z

Since ,the page rank of the page "X" depends upon the pagerank of page "Z"., page rank of the page "Y" depends upon the pagerank of page "X" and page rank of page "Z" depends on pageranks of page "X" and page "Y" .So, it is clear that we cannot calculate the pagerank of any page independently.So here we make an assumption .We assume that initially

PR(Z)= 0.20.

PROPOSED WEIGHTED PAGERANK ALGORITHM :

Now, we discuss the implmentation of our proposed weighted pagerank algorithm with respect of following figure.In this figure ,numbers on edges indicate the number of visits on the links by users. We know ,the page rank of a page "u" is calculated using the following formulae:

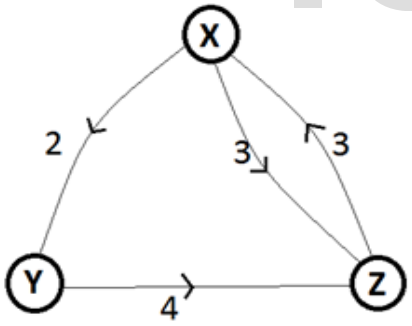


Fig: Proposed Weighted Pagerank Algorithm

$$wpr(u) = (1-d) + d \sum_{v \in B(u)} \left(\frac{l(uv)}{tl(v)} * out(vu) * wpr(v)\right)$$

So, pagerank of page "x" is given as:

wpr(x)= (1-d)+ d ((l(xz)/tl(z))*out(zx)*wpr(z))

Pagerank of page "y" is given as:

wpr(y)= (1-d)+ d ((l(yx)/tl(x))*out(xy)*wpr(x))

Pagerank of page "z" is given as:

$$wpr(z) = (1-d) + d \left(\frac{l(zx)}{tl(x)} * out(xz) * wpr(x) + \frac{l(zy)}{tl(y)} * out(yz) * wpr(y)\right)$$

From the above diagram we have

o(x)=2 // two edges emanating from x

o(y)=1 //one edge emanating from y

o(z)=1 //one edge emanating from z

Also,

out(zx) = o(x) / o(x)+o(y) = 2/3

out(xy) = o(y) / o(z) = 1

out(xz) = o(z) / o(z) = 1

out(yz) = o(z) / o(x) = 1/2

Also,

l(xz) / tl(z) = 3/3 = 1

l(zx) / tl(x) = 3/5

l(yx) / tl(x) = 2/5

l(zy) / tl(y) = 4/4 = 1

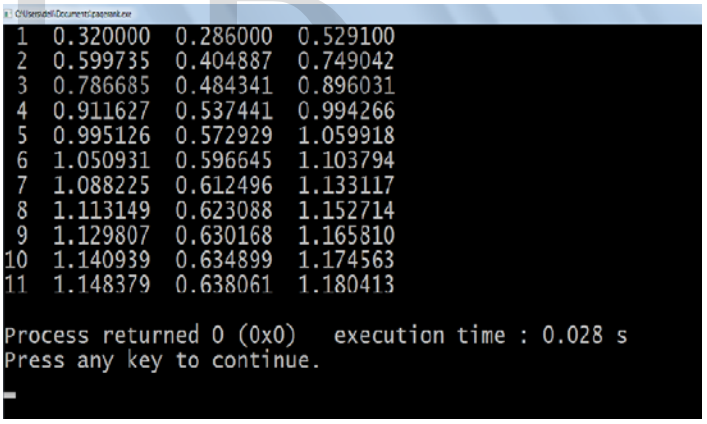
Since ,the page rank of the page "X" depends upon the pagerank of page "Z"., page rank of the page "Y" depends upon the pagerank of page "X" and page rank of page "Z" depends on pageranks of page "X" and page "Y" .So, it is clear that we cannot calculate the pagerank of any page independently.So here we make an assumption .We assume that initially

wpr(z)= 1.000000.

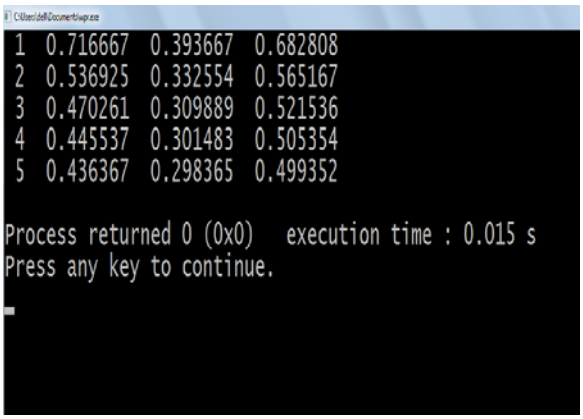
8 OBSERVATIONS

8.1 Considering d=0.85

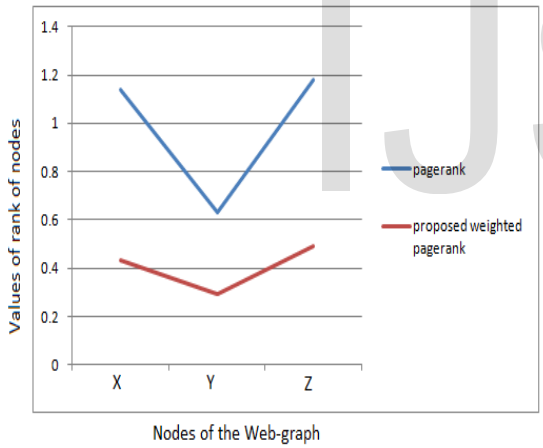
On executing the source code of the pagerank algorithm ,it gives the output as shown in the following snapshot



On executing the source of the proposed weighted pagerank algorithm ,it gives the output as shown in the following snapshot:

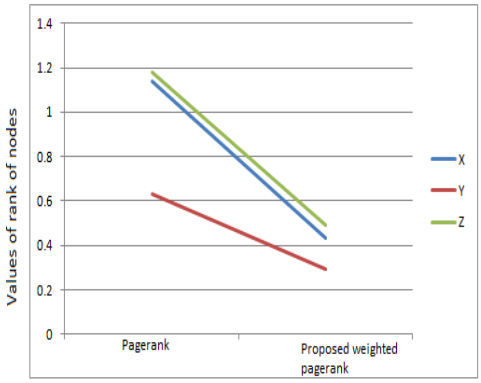


Having obtained the output on executing the program for the pagerank algorithm and the proposed weighted pagerank algorithm respectively, we observe that the number of computation steps was considerably reduced, steps which were involved in the computation of pagerank values iteratively. In case of pagerank algorithm it took 11 steps for it to come to an feasible approximate value , whereas it only took 5 steps for the proposed weighted pagerank algorithm to obtain the values within similar approximation.



Nodes of the Web-graph

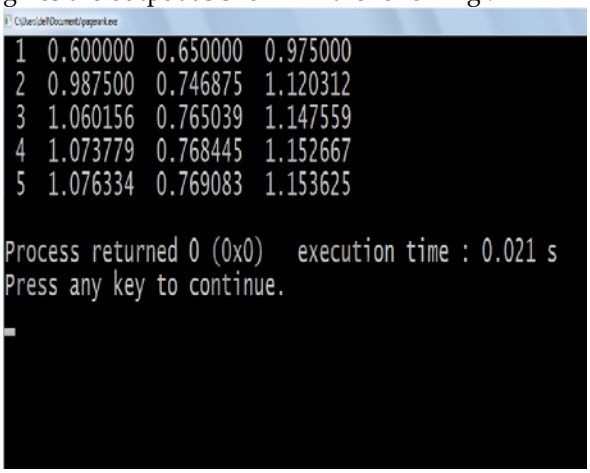
Also, this reduction was observed without any change in the result. The above graph is drawn for the three nodes shown on a horizontal line , with the values shown on the vertical line. We observe that the order of the rank of pages remain unchanged i.e they are relatively same ,only their absolute values are changed. Also,the below graph can be observed as ,showing the behaviour of different nodes (webpages) on the two algorithms. The following graph shows the two algorithms on the horizontal axis and the values of rank of nodes on the vertical axis.



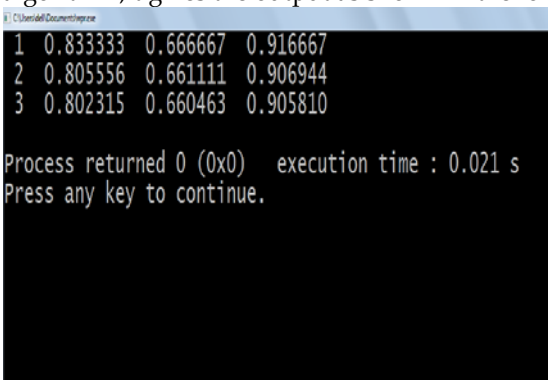
Comparing the two algorithms

8.2 Considering d=0.50

On executing the source code of the pagerank algorithm ,it gives the output as shown in the following :



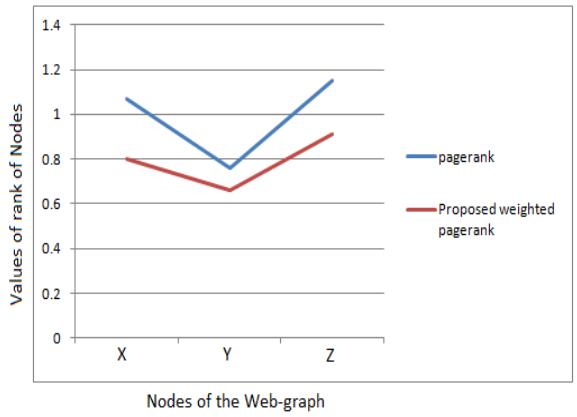
On executing the source of the proposed weighted pagerank algorithm ,it gives the output as shown in the following :



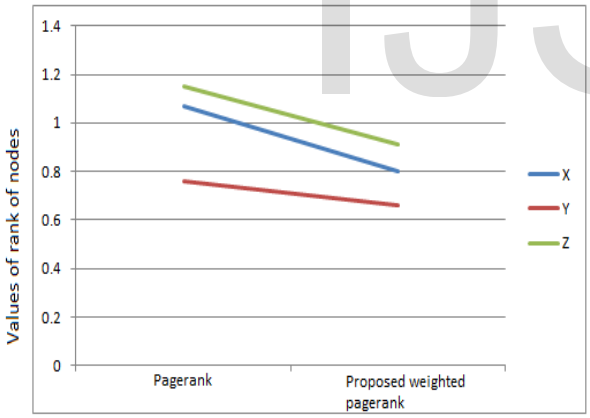
Having obtained the output on executing the program for the pagerank algorithm and the proposed weighted pagerank algorithm respectively, we again observe that the number of computation steps was considerably reduced, steps which were involved in the computation of pagerank values iteratively. In case of pagerank algorithm it took 5 steps for it to come to an feasible approximate value , whereas it only took 3 steps for the proposed weighted pagerank algorithm to

obtain the values within similar approximation.

Also, this reduction was observed without any change in the result. Below is a graph drawn for the three nodes shown on a horizontal line , with the values shown on the vertical line. We observe that the order of the rank of pages remain unchanged i.e they are relatively same ,only their absolute values are changed.

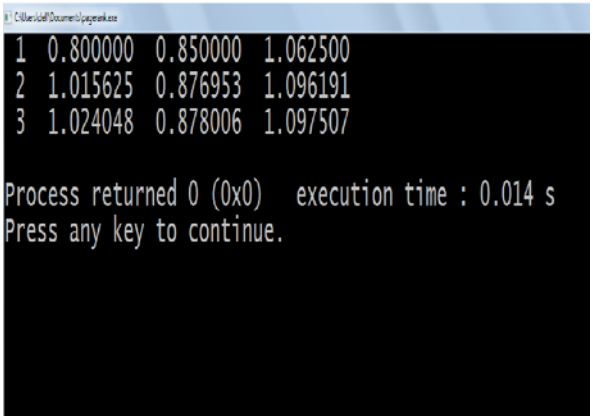


Also, the above graph can be observed as ,showing the behaviour of different nodes (webpages) on the two algorithms. The following graph shows the two algorithms on the horizontal axis and the values of rank of nodes on the vertical axis.

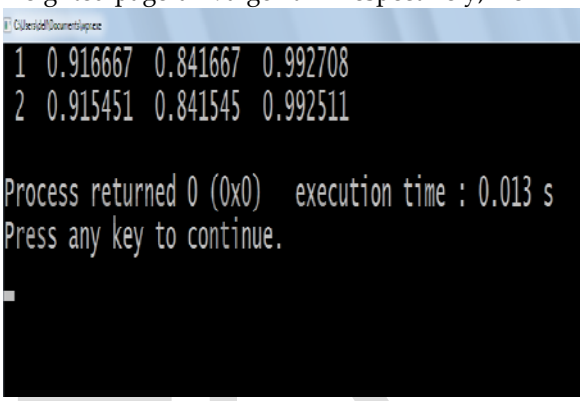


8.3 Considering d=0.25

On executing the source code of the pagerank algorithm ,it gives the output as shown in the following :

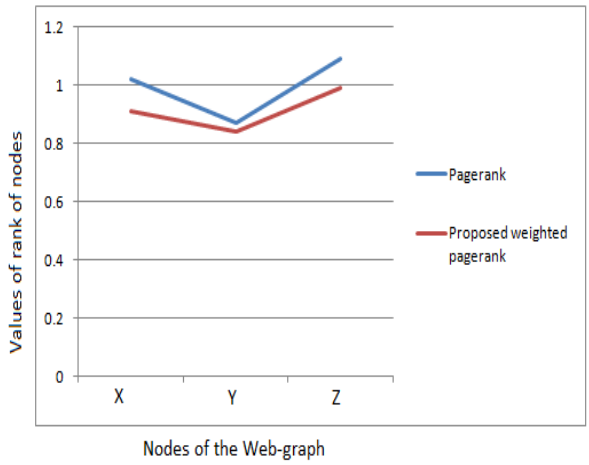


On executing the source of the proposed weighted pagerank algorithm, it gives the output as shown in the following snapshot. Having obtained the output on executing the program for the pagerank algorithm and the proposed weighted pagerank algorithm respectively, we

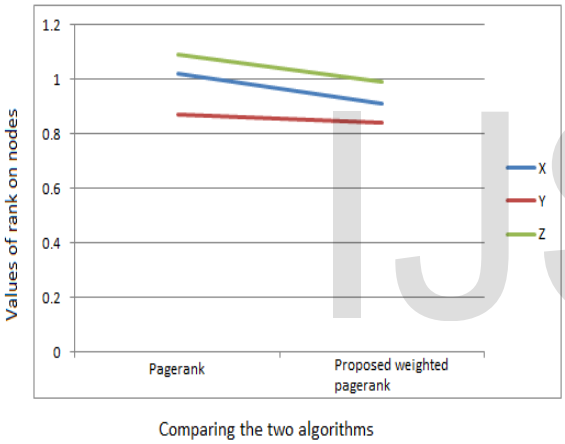


again observe that the number of computation steps was considerably reduced, steps which were involved in the computation of pagerank values iteratively. In case of pagerank algorithm it took 3 steps for it to come to an feasible approximate value , whereas it only took 2 steps for the proposed weighted pagerank algorithm to obtain the values within similar approximation.

Also, this reduction was observed without any change in the result. Below is a graph drawn for the three nodes shown on a horizontal line , with the values shown on the vertical line. We observe that the order of the rank of pages remain unchanged i.e they are relatively same ,only their absolute values are changed.



Also, the above graph can be observed as ,showing the behaviour of different nodes (webpages) on the two algorithms. The following graph shows the two algorithms on the horizontal axis and the values of rank of nodes on the vertical axis.



Thus, from the above ouputs and graphs obtained using d=0.85 , d=0.50 and d=0.25 , we observe that as we reduce the value of dampening factor "d" , the difference in number of computation steps involved decreases. At d=0.25, the number of computation steps involved was almost same in both cases. This can also be justified from the fact that as "d" decreases ,the fixed component of page rank value increases and the algorithm based component decreases. Thus, at lower "d" , there is no significant difference between the two algorithm

9 Future Scope

In the above section, we modified weighted page ranking algorithm to generate a better page ranking algorithm. This modified algorithm not only consider link structure but it also includes the users visit on a particular page. This modified algorithm provides more relevant results. The proposed algorithm, however, uses only link visit information from a user. 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.

Acknowledgment

The authors wish to thank Mr. Sanjay Kumar (CSE Dept. , NIT Jamshedpur).

References

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

[2] 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.

[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] Neelam Tyagi, Simple Sharma " Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[5] 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.

[6] Ian H. Witten , "How the Dragons Work: Searching in a Web" Department of Computer Science,University of Waikato, New Zealand,International Workshop On Research Issues in Digital Libraries(IWEIDL 2006)