JASEM ISSN 1119-8362
All rights reserved

Full-text Available Online at
www.ajol.info and
www.bioline.org.br/ja

J. Appl. Sci. Environ. Manage. *December 2013*
Vol. 17 (4) 551-558

The Illusion in the Presentation of the Rank of a Web Page with Dangling Links

[*1]FELIX UKPAI OGBAN; PRINCE OGHENEKARO ASAGBA (PH.D.); OLUMIDE OWOLABI (PH.D.)

[1]Department of Mathematics, Statistics, & Computer Science, Faculty of Science,
University of Calabar, Nigeria
Email: felix.ogban@gmail.com
[2]Department of Computer Science, Faculty of Physical & Information Technology, University of Port Harcourt, Nigeria
Email: pasagba@yahoo.com

[3]Computer Center, University of Abuja, Nigeria

Keywords: Hyperlinked, Page Rank algorithm, Page Rank calculation, Google Toolbar, in-bound and out-bound links, web page.

ABSTRACT: The hyperlinked display list of search results from any given search engine is an illusion of the real order in priority if its position is based on a pageRank computation. This is because, several factors governs the pageRank computation. Besides the fact that different search engines fashions their ranking model, it is an established fact that, the **hub** and **authority** factor must be considered. This paper considered the effects of an in-bound (authority) and out-bound (hub) links on the rank of a page. **Hyperlink-Induced Topic Search (HITS)** (also known as **hubs and authorities**) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg (1999). It was a precursor to PageRank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. A good hub page is one that points to many good authorities; a good authority page is one that is pointed to by many good hub pages. We focused on the Google's Toolbar with regard to pages given a certain toolbar Page Rank, but with an inbound link from a page that has a toolbar Page Rank, which is higher by one. This is done to alleviate the effect of the removal from the database(s), pages with no outbound links as proposed by Page and Brin, and applied by Google for the normalization of the dangling links. We considered, six (6) Web site clips scenarios to show the effects of inbound and outbound links, vis a vis the number of Pages in the Web, and the influence of the Damping Factor**.** We observed that the linkage of sets of dangling sites/pages (PDF, MS-word infested pages) and the application of a new ranking model for them is better in smoothening, the hub and authority complimentary effects in page ranking. We gave recommendations on how search engines and crawlers could weight and produce a better ranking of pages for users. Thus, users can within the first few search results, get to their expected search results. © JASEM

Some people may have the idea that generating millions of pages is a good way to produce Page Rank and improve ranking of their website after studying the Page Rank algorithm. Theoretically, this should work, assuming that an appropriate linking structure is chosen. However, it does not practically work that way. **Google** changed their algorithm several years ago. One of the changes prevents the generation of Page Rank by weighted suggestions from page designers (self-links) (Googlewebmaster, 2012). Moreover in practice, you even need Page Rank to get larger websites completely crawled. Again, Google made some minor changes of the algorithm, that is, they changed the value of the damping factor which originally was d=0.85 to 0.50

Another aspect is the question: which links are counted for Page Rank calculation by Google? If page **A** is linking two times to page **B** and one time to page **C**, then there are different possibilities how Page Rank is split between page **B** and **C**. The current implementation of the Page Rank algorithm is ignoring multiple links, that is, in the example above, the transferred Page Rank is split 1 by 1 between page **B** and page **C**. Self-links are another example of ambiguousness. Currently, Google is counting self-links for Page Rank calculation. However, links with the attribute ***rel="nofollow"*** are ignored just as robots.txt file contents are ignored during crawling.

Currently, there are just two ways to get independent information about the Page Rank: (1) the Google toolbar, and (2) Google directory. Of course, there are

Corresponding author Email: pasagba@yahoo.com

people offering tools which display Page Rank information directly within the Browser (via HTML). However, normally they are using just toolbar information for these services (GoogleToolbarGroup, 2013). Google's toolbar is using a logarithmic scale in the range 0 to 10. Due to the logarithmic scale, an increase of 1 in the toolbar Page Rank corresponds to an increase of the real Page Rank by the coefficient b the logarithmic base. Most likely Google calibrated/normalized the scale in such a way that the page with highest Page Rank is set to exactly 11. Google's directory was using a logarithmic scale in the range 0 to 7 until early 2008. Due to the different scale, one could get additional, more accurate information about the Page Rank, that is, pages with the same toolbar Page Rank might show a different directory value. Additionally, one could use the fact that pages within the directory are ordered by Page Rank. Of course, one could get this information just for pages which are listed in the site link pages (since Google uses these data). Nevertheless, problems occurred due to different update times of the toolbar and the directory. It seems that normally the directory values were more up to date. In the meantime, Google is using the toolbar values (0 - 10) also for the directory (Googlewebmaster, 2012). Before now, there were a lot of attempts to manipulate Page Rank shown in the toolbar. This can be done by redirecting a page to another one with high Page Rank. Google merges these pages and shows the Page Rank of the target page also for the redirecting page. If the redirection is replaced by new content the toolbar is showing the false Page Rank for a while (GoogleToolbarGroup, 2013). Also, sometimes pages are cloaked and the redirection is only visible for the search engines while visitors see the normal content. False Page Rank can be detected by examining cached and archived version of the page, reviewing the back links, checking if Page Rank is passed to other pages and using the 'info'-command.

Page Rank calculators compute numerically the Page Rank for some chosen linking structure. Page Rank calculators can be used to understand the basic mechanisms of the calculation (Abiteboul et al, 2003). Of course, one can examine just small web sites. Also, external incoming and outgoing links are often not considered. Also, neither Google's modifications are taken into account nor can one examine complex linking structure. Therefore, calculators are good for educational use, but not for practical guidance when developing linking structures.

There are several web sites on the Internet claiming to predict future Page Rank for a page. Of course, nobody has the same linking structure of the Internet as Google to compute Page Rank. Therefore, the prediction must be based on some data (Lawrence and Giles, 2000). The number of incoming links (which is extremely inaccurate) or the incoming links, plus the corresponding toolbar Page Rank is taken for the prediction. However, even in the second case there are numerous unknowns: There is a problem if the back-links are taken from Google because Google does not display all of them. There is also a problem if the back-links are taken from another search engine because it is not clear if Google takes the same links into account. Above all, during links crawling, other factors must be considered such as:

i. It is impossible to check if Page Rank is passed to other pages.
ii. The toolbar is showing only an integer (on a logarithmic scale). PR6 might be a 6.0 or 6.99 on the toolbar scale. Therefore, one has to take an average for calculation. iii. The values shown in toolbar for dynamic URLs are incorrect. iv. The currently shown values of the toolbar are taken as input. However, these values changes during the next update. v. The logarithmic base is part of the calculation for the prediction. However, most of the people are using completely wrong values in the range between 5 and 8. vi. The number of outgoing links must be replaced by an average value. vii. The toolbar scale is not fixed.

All these errors add up and make a reasonable prediction impossible. Most likely the hit rate is not higher than a simple prediction of the form 'Page Rank would not change' or taking the maximum Page Rank of a linking page minus one.

DEFINITION OF THE PROBLEM

Users of the Google Toolbar often notice that pages with a certain toolbar Page Rank have an inbound link from a page with a toolbar Page Rank which is higher by one. Some take this observation to doubt the validity of the Page Rank algorithm for the actual ranking methods of the Google search engine.

However, the number of outbound links on the linking page thwarts the effect of the logarithmical basis, because the Page Rank propagation from one page to another is divided by the number of outbound links on the linking page (Abiteboul et al; 2003). The Page Rank benefit by a link is higher than Page Rank algorithm's term **d(PR(Ti)/C(Ti))**. The reason is that the Page Rank benefit for one page is further distributed to other pages within the site. If those pages link back as it usually happens, the Page Rank benefit for the page which initially received the link. If we assume that at a high damping factor the logarithmical basis for Page Rank scaling is 6 and a page receives a Page Rank benefit which is twice as high as the Page Rank of the linking page divided by the number of its outbound links, the linking page could have at least 12 outbound links so that the Toolbar Page Rank of the page receiving the link is still at most one lower than the toolbar Page Rank of the linking page. But this is not so in the case of Google Toolbar display thus misleading to users. A number of 12 outbound links admittedly seems relatively small. But normally, if a page has an external inbound link, this is not the only one for that page. Most likely other pages link to that page and propagate Page Rank to it. And if there are examples where a page receives a single link from another page and the Page Ranks of both pages comply the PageRank-1 rule although the linking page has many outbound links, this is first of all an indication for the linking page's toolbar Page Rank being at the upper end of its scale. The linking page could be a "high" 5 and the page receiving the link could be a "low" 4. In this way, the linking page could have up to 72 outbound links. This number rises accordingly if we assume a higher logarithmical basis for the scaling of Toolbar Page Rank. But this is not the case in reality, thus the need to probe the situation.

PageRank-1 rule proves the fundamental principle of PageRank. Web pages are important themselves if other important web pages link to them. That is a page is important, if at least one other important page, links to it. This produces a chain of important pages.

Page Ranking Factors And Linkages: For the rank of a page to be gotten, several factors must be involved. While some reduces the ranking, some increases it. The structure of the site and the linkage plays a very important role in this growth and retardation of the value of a page's rank. These factors are treated one after the other to ascertain the Illusion in the Presentation of the Rank of a Web Page with respect to their links.

The Effect of Inbound Links: Each additional inbound link for a web page always increases that page's Page Rank. Taking a look at the Page Rank algorithm, this is given by:

$$PR(A)=(1-d)+d(PR(T1)/C(T1)+\ldots+PR(Tn)/C(Tn)) \quad (1)$$

One may assume that an additional inbound link from page **X** increases the Page Rank of page **A** by:

$d \times PR(X) / C(X)$ (2)

where PR(X) is the Page Rank of page **X**, C(X) is the total number of its outbound links. But page **A** usually links to other pages itself. Thus, these pages get a Page Rank benefit also. If these pages link back to page **A,** it will have an even higher Page Rank benefit from its additional inbound link. Figure 1 shows Web site clip 1.The single effects of additional inbound links shall be illustrated by an example.

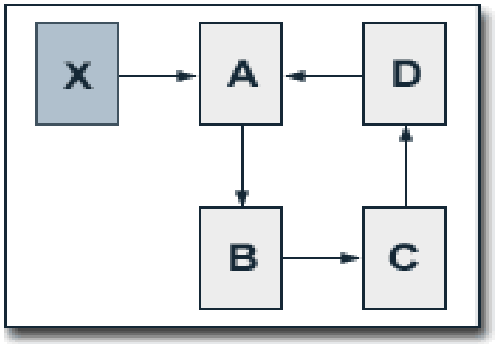

Fig. 1: Web site clip 1

In Figure 1, we regard a website consisting of four pages **A, B, C** and **D**, which are linked to each other in circle. Without external inbound links to one of these pages, each of them obviously has a Page Rank of 1. We now add a page X to our example, for which we presume a constant Page rank PR(X) of 10. Page X links to page A by its only outbound link. Setting the damping factor *d* to 0.5, we got the following computations/equations for the Page Rank values of the single pages of our site:

PR(A) = 0.5 + 0.5 (PR(X) + PR(D)) = 5.5 + 0.5 PR(D)
PR(B) = 0.5 + 0.5 PR(A)
PR(C) = 0.5 + 0.5 PR(B)
PR(D) = 0.5 + 0.5 PR(C)

Since the total number of outbound links for each page is one, the outbound links do not need to be considered in the equations. Solving them gives us the following Page Rank values:

PR(A) = 19/3 = 6.33
PR(B) = 11/3 = 3.67
PR(C) = 7/3 = 2.33
PR(D) = 5/3 = 1.67

We noticed that the initial effect of the additional inbound link of page A, which was given by:

*d * PR(X) / C(X) = 0.5 * 10 / 1 = 5*

is passed on by the links on our site.

The Influence of the Damping Factor: The degree of Page Rank propagation from one page to another by a link is primarily determined by the damping factor d. If we set *d* to 0.75, we got the following equations from the previous ones:

PR(A) = 0.25 + 0.75 (PR(X) + PR(D)) = 7.75 + 0.75 PR(D)
PR(B) = 0.25 + 0.75 PR(A)
PR(C) = 0.25 + 0.75 PR(B)
PR(D) = 0.25 + 0.75 PR(C)

Solving these equations gives us the following Page Rank values:

PR(A) = 419/35 = 11.97
PR(B) = 323/35 = 9.23
PR(C) = 251/35 = 7.17
PR(D) = 197/35 = 5.63

First of all, we noticed that there is a significantly higher initial effect of additional inbound link for page A which is given by: $d * PR(X) / C(X) = 0.75 * 10 / 1 = 7.5$

This initial effect is then propagated even stronger by the links on our site. In this way, the Page Rank of page A is almost twice as high at a damping factor of 0.75, than it was at a damping factor of 0.5. At a damping factor of 0.5 the Page Rank of page A is almost four times superior to the Page Rank of page D, while at a damping factor of 0.75 it is only a little more than twice as high.

So, the higher the damping factor, the larger is the effect of an additional inbound link for the Page Rank of the page that receives the link and the more evenly distributes Page Rank over the other pages of a site (Page and Brin; 1998).

The Actual Effect of Additional Inbound Links: At a damping factor of 0.5, the accumulated Page Rank of all pages of our site is given by: $PR(A) + PR(B) + PR(C) + PR(D) = 14$

Hence, by a page with a Page Rank of 10 linking to one page of our example site by its only outbound link, the accumulated Page Rank of all pages of the site is increased by 10. Before adding the link, each page has had a Page Rank of 1. At a damping factor of 0.75, the accumulated Page Rank of all pages of the site is given by:

$PR(A) + PR(B) + PR(C) + PR(D) = 34$ (3)

This time the accumulated Page Rank increases by 30. The accumulated Page Rank of all pages of a site always increases by: $(d / (1-d)) * (PR(X) / C(X)$ (4) where X is a page additionally linking to one page of the site, PR(X) is its Page Rank and, C(X) its number of outbound links. The formula presented above is only valid, if the additional link points to a page within a closed system of pages, for instance, a website without outbound links to other sites.

As far as the website has links pointing to external pages, the surplus for the site itself diminishes accordingly, because a part of the additional Page Rank is propagated to external pages (Cho and Garcia-Molina; 2003). The justification of the above formula of Equation (4) is given by Raph Levien (2009) and it is based on the Random Surfer Model. The walk length of the random surfer is an exponential distribution with a mean of (d/(1-d)). When the random surfer follows a link to a closed system of web pages, she visits on average (d/(1-d)) pages within that closed system. So, this much more Page Rank of the linking page - weighted by the number of its outbound links - is distributed to the closed system (Ipeirotis, et al, 2005).

For the actual Page Rank calculations at Google, Lawrence Page und Sergey Brin (1998), claimed to usually set the damping factor d to 0.85. Thereby, the boost for a closed system of web pages by an additional link from page X is given by:

$(0.85 / 0.15) * (PR(X) / C(X)) = 5.67 * (PR(X) / C(X))$ (5)

So, inbound links have a far larger effect than one may assume.

The Effect of Outbound Links: Since Page Rank is based on the linking structure of the whole web, it is inescapable that if the inbound links of a page influence its Page Rank, its outbound links do also have some impact. Figure 2 shows web site clip 2.

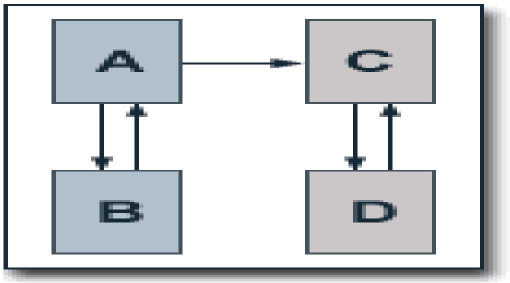

Fig. 2: Web site clip 2

In Figure 2, we regard a web consisting of two websites, each having two web pages. One site consists of pages A and B, the other consists of pages C and D. Initially, both pages of each site solely link to each other. It is obvious that each page then has a Page Rank of one as proposed by (Brin and page, 1998).

Now, we add a link which points from page A to page C. At a damping factor of 0.75, we therefore obtain the following equations for the single pages' Page Rank values:

$PR(A) = 0.25 + 0.75\ PR(B)$

$PR(B) = 0.25 + 0.375\ PR(A)$

$PR(C) = 0.25 + 0.75\ PR(D) + 0.375\ PR(A)$

$PR(D) = 0.25 + 0.75\ PR(C)$

Solving the equations gives us the following Page Rank values for the first site:

$PR(A) = 14/23$

$PR(B) = 11/23$

We therefore get an accumulated Page Rank of 25/23 for the first site. The Page Rank values of the second site are given by

$PR(C) = 35/23$

$PR(D) = 32/23$

So, the accumulated Page Rank of the second site is 67/23. The total Page Rank for both sites is 92/23 = 4. Hence, adding a link has no effect on the total Page Rank of the web. Additionally, the Page Rank benefit for one site equals the Page Rank loss of the other.

The Actual Effect of Outbound Links: As it has already been shown, the Page Rank benefit for a closed system of web pages by an additional inbound link is given by: (d / (1-d)) * (PR(X) / C(X)), (6) where X is the linking page, PR(X) is its Page Rank, and C(X) is the number of its outbound links. Hence, this value also represents the Page Rank loss of a formerly closed system of web pages, when a page X within this system of pages now points by a link to an external page.

The validity of the above formula requires that the page which receives the link from the formerly closed system of pages does not link back to that system, since it otherwise gains back some of the lost Page Rank. Of course, this effect may also occur when not the page that receives the link from the formerly closed system of pages links back directly, but another page which has an inbound link from that page. Indeed, this effect may be disregarded because of the damping factor, if there are enough, other web pages in-between the link-recursion (Cho, et al, 2003). The validity of the formula also requires that the linking site has no other external outbound links. If it has other external outbound links, the loss of Page Rank of the regarded site diminishes and the pages already receiving a link from that page lose Page Rank accordingly. Even if the actual Page Rank values for the pages of an existing web site were known, it would not be possible to calculate to which extend an added outbound link diminishes the Page Rank loss of the site, since the above presented formula regards the status after adding the link.

Intuitive Justification of the Effect of Outbound ***Links:*** The intuitive justification for the loss of Page Rank by an additional external outbound link according to the Random Surfer Model, is that, by adding an external outbound link to one page, the surfer will less likely follow an internal link on that page. So, the probability for the surfer reaching other pages within a site diminishes. If those other pages of the site have links back to the page to which the external outbound link has been added, also this page's Page Rank will deplete. Lastly, relevant outbound links do constitute the quality of a web page and a webmaster who points to other pages integrates their content in some way into his own site (Ipeirotis et al; 2005). We can conclude that external outbound links diminish the totalized Page Rank of a site and probably also the Page Rank of each single page of a site. But, since links between web sites are the fundamental of Page Rank and indispensable for its functioning, there is the possibility that outbound links have positive effects within other parts of Google's ranking criteria.

Dangling Links: An important aspect of outbound links is the lack of them on web pages. When a web page has no outbound links, its Page Rank cannot be distributed to other pages. Lawrence Page and Sergey Brin characterized links to those pages as dangling links (Brin and Page; 1998). Figure 3 shows web site clip 3.

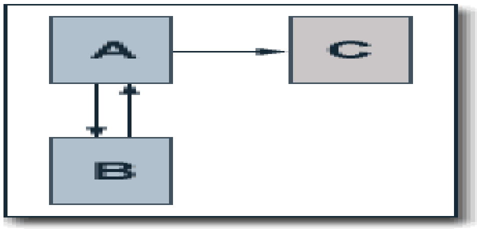

Fig. 3: Web site clip 3

In Figure 3, the effect of dangling links shall be illustrated by a small example website. We considered a site consisting of three pages **A, B** and **C**. In this case, the pages **A** and **B** link to each other. Additionally, page **A** links to page **C**. Page **C** itself has no outbound links to other pages. At a damping factor of 0.75, we obtained the following equations for the single pages' Page Rank values:

$PR(A) = 0.25 + 0.75\ PR(B)$
$PR(B) = 0.25 + 0.375\ PR(A)$
$PR(C) = 0.25 + 0.375\ PR(A)$

Solving the equations gives us the following Page Rank values:

$PR(A) = 14/23$
$PR(B) = 11/23$
$PR(C) = 11/23$

So, the accumulated Page Rank of all three pages is 36/23 which is just over half the value that we could have expected if page A had links to one of the other pages. According to Page and Brin (1998), the number of dangling links in Google's index is fairly high. The reason therefore is that many linked pages are not indexed by Google, for example because indexing is disallowed by a robots.txt file (Abiteboul, et al, 2003). Additionally, Google meanwhile indexes several file types and not HTML only. PDF or Word files do not really have outbound links and, hence, dangling links could have major impacts on Page Rank. Figure 4 shows web site clip 4.

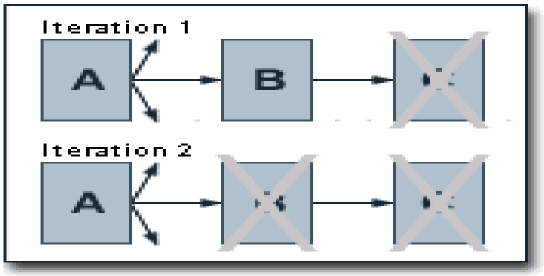

Fig. 4: Web site clip 4

In order to prevent Page Rank from the negative effects of dangling links, pages without outbound links have to be removed from the database until the Page Rank values are computed. According to Page and Brin, (1998), the number of outbound links on pages with dangling links is thereby normalized. As shown in Figure 4, removing one page can cause new dangling links and, hence, removing pages has to be an iterative process. After the Page Rank calculation is finished, Page Rank can be assigned to the formerly removed pages based on the Page Rank algorithm. Therefore several iterations are needed for removing the pages. Regarding our illustration, page **C** could be processed before page **B**. At that point, page **B** has no Page Rank yet and, so, page **C** will not receive any either. Then, page **B** receives Page Rank from page **A** and during the second iteration, also page **C** gets its Page Rank. In Figure4 - website for dangling links, removing page C from the database results in page A and B each having a Page Rank of 1. After the calculations, page C is assigned a Page Rank of 0.25 + 0.375 PR(A) = 0.625. So, the accumulated Page Rank does not equal the number of pages, but at least all pages which have outbound links are not harmed from the dangling links problem. By removing dangling links from the database, they do not have any negative effects on the Page Rank of the rest of the web. Since PDF files are dangling links, links to PDF files do not diminish the Page Rank of the linking page or site. So, PDF files can be a good means of search engine optimization for Google if not in a nofollow link.

Effect of the Number of Pages in the Web

Since the accumulated Page Rank of all pages of the web equals the total number of web pages, it follows directly that an additional web page increases the added up Page Rank for all pages of the web by one (Lawrence and Giles; 2000). But far more interesting than the effect on the added up Page Rank of the web is the impact of additional pages on the Page Rank of actual websites. To illustrate the effects of additional web pages, we considered a hierarchically structured web site, consisting of three pages: **A, B** and **C**, which are joined by an additional page **D** on the hierarchically lower level of the site as shown in Figure 5 (web site clip 5).

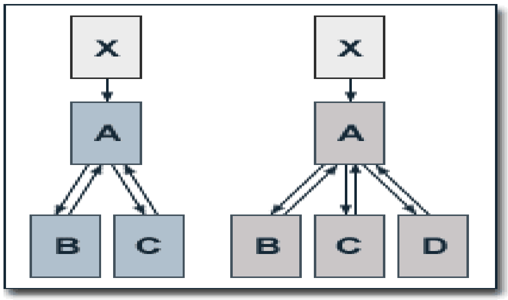

Fig. 5: Web site clip 5

The site has no outbound links. A link from page **X** which has no other outbound links and a Page Rank of 10 points to page **A**. At a damping factor d of 0.75, the equations for the single pages' Page Rank values before adding page **D** are given by:

$PR(A) = 0.25 + 0.75\,(10 + PR(B) + PR(C))$
$PR(B) = PR(C) = 0.25 + 0.75\,(PR(A) / 2)$

Solving the equations gives us the following Page Rank values:

$PR(A) = 260/14$
$PR(B) = 101/14$
$PR(C) = 101/14$

After adding page **D**, the equations for the pages' Page Rank values are given by

$PR(A) = 0.25 + 0.75\,(10 + PR(B) + PR(C) + PR(D))$
$PR(B) = PR(C) = PR(D) = 0.25 + 0.75\,(PR(A) / 3)$

Solving these equations gives us the following Page Rank values:

$PR(A) = 266/14$
$PR(B) = 70/14$
$PR(C) = 70/14$
$PR(D) = 70/14$

Figure 5 has no outbound links, but after adding page **D**, the accumulated Page Rank of all pages increases by one from 33 to 34. Further, the Page Rank of page **A** rises marginally. In contrast, the Page Rank of pages **B** and **C** depletes substantially. The Reduction of Page Rank by Additional Pages by adding pages to a hierarchically structured websites, the consequences for the already existing pages are non-uniform. The consequence(s) for websites with a different structure is presented. Figure 6 shows web site clip 6.

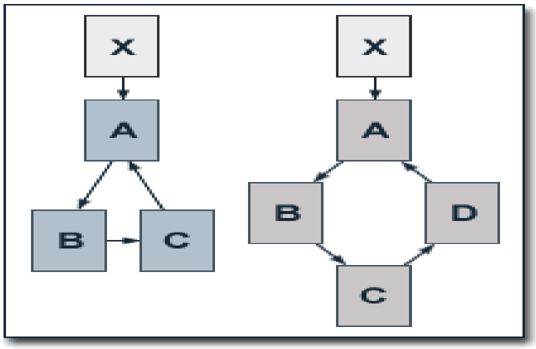

Fig. 6: Web site clip 6

In Figure 6, we examined a website consisting of three pages **A, B** and **C** which are linked to each other in circle and the pages are then joined by page **D** which fits into the circular linking structure. The regarded site has no outbound links. Again, a link from page **X** which has no other outbound links and a Page Rank of 10 points to page **A**. At a damping factor d of 0.75, the equations for the single pages' Page Rank values before adding page **D** are given by:

$PR(A) = 0.25 + 0.75\,(10 + PR(C))$
$PR(B) = 0.25 + 0.75 * PR(A)$
$PR(C) = 0.25 + 0.75 * PR(B)$

Solving the equations gives us the following Page Rank values:

$PR(A) = 517/37 = 13.97$
$PR(B) = 397/37 = 10.73$
$PR(C) = 307/37 = 8.30$

After adding page **D**, the equations for the pages' Page Rank values are given by

$PR(A) = 0.25 + 0.75 (10 + PR(D))$
$PR(B) = 0.25 + 0.75 * PR(A)$
$PR(C) = 0.25 + 0.75 * PR(B)$
$PR(D) = 0.25 + 0.75 * PR(C)$

Solving these equations gives us the following Page Rank values:

$PR(A) = 419/35 = 11.97$
$PR(B) = 323/35 = 9.23$
$PR(C) = 251/35 = 7.17$
$PR(D) = 197/35 = 5.63$

Again, after adding page **D**, the accumulated Page Rank of all pages increases by one from 33 to 34. But now, any of the pages which already existed before page **D** was added lose Page Rank. The more uniform Page Rank is distributed by the links within a site, the more likely will this effect occur.

DISCUSSION OF RESULTS

In web site clip 1, we observed that the initial effect of the additional inbound link of page A, which was given by: d * PR(X) / C(X) is passed on by the links on our site.

In web site clip 2, we observed that adding a link (Outbound) has no effect on the total Page Rank of the web. Additionally, the Page Rank benefit for one site equals the Page Rank loss of the other all things being equal. As it has already been shown, the Page Rank benefit for a closed system of web pages by an additional inbound link is given by:
(d / (1-d)) * (PR(X) / C(X)), where X is the linking page, PR(X) is its Page Rank, and C(X) is the number of its outbound links. Hence, this value also represents the Page Rank loss of a formerly closed system of web pages, when a page X within this system of pages now points by a link to an external page.

In web site clip 3, we observed that According to Page and Brin (1998), the number of dangling links in Google's index is fairly high. The reason therefore is that many linked pages are not indexed by Google, for example because indexing is disallowed by a robots.txt file. Additionally, Google meanwhile, indexes several file types and not HTML only. PDF or Word files do not really have outbound links and, hence, dangling links could have major impacts on the Rank of pages.

In web site clip 4, we observed that by removing dangling links from the database, they do not have any negative effects on the Page Rank of the rest of the web. Since PDF files are dangling links, links to PDF files do not diminish the Page Rank of the linking page or site. So, PDF files can be a good means of search engine optimization for Google if not in a nofollow link.

In web site clip 5, we observed that Since the accumulated Page Rank of all pages of the web equals the total number of web pages, it follows directly that an additional web page increases the added up Page Rank for all pages of the web by one (Lawrence and Giles; 2000). But far more interesting than the effect on the added up Page Rank of the web is the impact of additional pages on the Page Rank of actual websites which is found not to be uniform. We found out that there is a reduction of Page Rank by Additional Pages by adding pages to a hierarchically structured websites, and that the consequences for the already existing pages are non-uniform to the pages.

In web site clip 6, we observed that adding pages to a site often reduces Page Rank for already existing pages, it becomes obvious that the Page Rank algorithm tends to privilege smaller web sites. Indeed, bigger web sites can counterbalance this effect by being more attractive for other webmasters to link to them, simply because they have more content (Najork and Wiener; 2001). Nevertheless, it is also possible to increase the Page Rank of existing pages by additional pages. Therefore, it has to be considered that as few Page Rank as possible is distributed to these additional pages.

Conclusion: Users of the Google Toolbar often notice that pages with a certain toolbar Page Rank have an inbound link from a page with a toolbar Page Rank which is higher by one. Some take this observation to doubt the validity of the Page Rank algorithm presented here for the actual ranking methods of the Google search engine. It was shown, however, that the PageRank-1 rule complies with the Page Rank algorithm. Basically, the PageRank-1 rule proves the fundamental principle of Page Rank. Web pages are important themselves if other important web pages link to them. It is not necessary for a page to have many inbound links to rank well. A single link from a high ranking page is sufficient.

To show the actual consistence of the PageRank-1 rule with the Page Rank algorithm, several factors have to be taken into consideration. First of all, the toolbar Page Rank is a logarithmically scaled version of real Page Rank values. If the Page Rank value of one page is one higher than the Page Rank value of another page in terms of Toolbar Page Rank, than its real Page Rank can at least be higher by an amount which equals the logarithmical basis for the escalation of Toolbar Page Rank. If the logarithmical basis for the escalation is 6 and the toolbar Page Rank of a linking Page is 5, then the real Page Rank of the page which receives the link can be at least 6 times smaller to make that page still get a toolbar Page Rank of 4. From our work, it is obvious that there are several illusions in the presentation or representation of the results of pages ranks from several search engines, except the web structures is normalized to a generic format such that the linked and the non-linked pages are clustered separately. Secondly, closed systems sites are not indexed along with systems sites that are open. Thirdly, the classification involved should be such that file formats and types are separated. In other words, classifying PDF, Doc, Docx, odt, etc files should be separated from that of HTML files since dangling links are involve in the earlier file types.

Recommendation: Putting the various findings together, we recommend that a page ranking system must be put in place when building search engines and crawlers. Several link related factors must be taken into consideration to guide against the losses and gains associated with inbound, outbound and dangling links. Secondly, the crawler should separate, during crawling, file types that cannot be linked to or from into a cluster whose ranking would be differently considered. With that, a level playing ground would be given to indexes, parsers, language miners, search engines and other meta crawlers to weight and produce a better ranking of pages for users. Thus, users can within the first few search results, get to their expected search results.

REFERENCES

Abiteboul, S., Preda, M., and Cobena, G. (2003). Adaptive on-line page importance computation. *In*: Proceedings of the twelfth international conference on World Wide Web (Budapest, Hungary: ACM Press): 280–290.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107–117.

Cho, J. and Garcia-Molina, H. (2003). Estimating frequency of change. ACM Transactions on Internet Technology, 3(3): 7 – 35

Googlewebmaster, (2012), More videos: http://www.youtube.com/GoogleWebmaster,
Webmaster Central Blog: http://googlewebmastercentral.blogspo, Webmaster Central: http://www.google.com/webmasters

GoogleToolbarGroup, (2013), Meta files indexing, ranking, and file indicator Group: http://www.GoogleToolbarGroup.com

Ipeirotis, P., Ntoulas, A., Cho, J., and Gravano, L. (2005), Modeling and managing content changes in text databases. In Proceedings of the 21st IEEE International Conference on Data Engineering, pages 606-617, Tokyo.

Kleinberg, J. M. (1999). "Authoritative sources in a hyperlinked environment". *Journal of the ACM* 46 (5): 604. doi:10.1145/324133.324140

Lawrence, S. and Giles, C. L. (2000), Accessibility of information on the web, Intelligence, 11(1) : 32–39.

Dunn, J. C. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics* 3: 32-57

Bezdek, J. C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York

Tariq R., T: (2002), Clustering: http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html

Hans-Joachim, M and Hizir, S (2007), Nonhierarchical Clustering: http://www.quantlet.com/mdstat/scripts/xag/html/xaghtmlframe149.html

Osmar, R. Z (2007), Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html

Marc-Najork and Wiener J. L. (2001) Breadth-first crawling yields high-quality pages. In Proceedings of the Tenth Conference on World Wide Web, , Elsevier Science, Hong Kong: 114–118.