Enhancing the Ranking of a Web Page in the Ocean of Data

Hitesh KUMAR SHARMA
University of Petroleum and Energy Studies, India
hkshitesh@gmail.com

In today's world, web is considered as ocean of data and information (like text, videos, multimedia etc.) consisting of millions and millions of web pages in which web pages are linked with each other like a tree. It is often argued that, especially considering the dynamic of the internet, too much time has passed since the scientific work on PageRank, as that it still could be the basis for the ranking methods of the Google search engine. There is no doubt that within the past years most likely many changes, adjustments and modifications regarding the ranking methods of Google have taken place, but PageRank was absolutely crucial for Google's success, so that at least the fundamental concept behind PageRank should still be constitutive. This paper describes the components which affects the ranking of the web pages and helps in increasing the popularity of web site. By adapting these factors website developers can increase their site's page rank and within the PageRank concept, considering the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them. The PageRank of a document is always determined recursively by the PageRank of other documents.

Keywords: *SEO, Search Engine, Page Rank, Inbound, Outbound, DBMS*

1 Introduction

PageRank was developed by Google founders Larry Page and Sergey Brin at Stanford. At the time that Page and Brin met, search engines typically linked to pages that had the highest keyword density, which meant people could game the system by repeating the same phrase over and over to attract higher search page results. The rapidly growing web graph contains several billion nodes, making graph-based computations very expensive. One of the best known web-graph computations is Page-Rank, an algorithm for determining the "importance" of Web pages. [7]. Page and Brin's theory is that the most important pages on the Internet are the pages with the most links leading to them. [1] PageRank thinks of links as votes, where a page linking to another page is casting a vote.

2. Page Rank

PageRank is the algorithm used by the Google search engine, originally formulated by Sergey Brin and Larry Page in their paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine".It is based on the premise, prevalent in the world of academia, that the importance of a research paper can be judged by the number of citations the paper has from other research papers. Brin and Page have simply transferred this premise to its web equivalent: the importance of a web page can be judged by the number of hyperlinks pointing to it from other web pages. [2] Now web graph has huge dimensions and is subject to dramatic updates in terms of nodes and links, therefore the PageRank assignment tends to became obsolete very soon [4]. It may look daunting to non-mathematicians, but the PageRank algorithm is in fact elegantly simple and is calculated as follows:

$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n))$

where:
PR(A) is the PageRank of a page A
$PR(T_1)$ is the PageRank of a page T_1
$C(T_1)$ is the number of outgoing links from the page T_1, d is a damping factor in the range $0 < d < 1$, usually set to 0.85.The PageRank of a web page is therefore calculated as a sum of the PageRanks of all pages linking to it (its *incoming links*), divided by the number of links on each of those pages (its *outgoing links*). From a search engine marketer's point of view, this means there are two ways in which PageRank can affect the position of your page on Google:

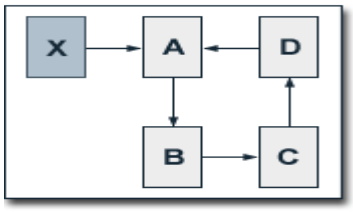
- *The number of incoming links.* Obviously the more of these the better. But there is another thing the algorithm tells us: no incoming link can have a negative effect on the PageRank of the page it points at. At worst it can simply have no effect at all.
- *The number of outgoing links on the page which points at your page.* The fewer of these the better. This is interesting: it means given two pages of equal PageRank linking to you, one with 5 outgoing links and the other with 10, you will get twice the increase in PageRank from the page with only 5 outgoing links. At this point we take a step back and ask ourselves just how important PageRank is to the position of your page in the Google search results. The next thing we can observe about the PageRank algorithm is that it has nothing whatsoever to do with relevance to the search terms queried. It is simply one single (admittedly important) part of the entire Google relevance ranking algorithm. Perhaps a good way to look at PageRank is as a multiplying factor, applied to the Google search results after all its other computations have been completed. The Google algorithm first calculates the relevance of pages in its index to the search terms, and then multiplies this relevance by the PageRank to produce a final list. The higher your PageRank therefore the higher up the results you will be, but there are still many other factors related to the positioning of words on the page which must be considered first.[2]

3. The Effect of Inbound Links

It has already been shown that each additional inbound link for a web page always increases that page's PageRank. Taking a look at the PageRank algorithm, which is given by:

$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

one may assume that an additional inbound link from page X increases the PageRank of page A by $d \times PR(X) / C(X)$ where PR(X) is the PageRank of page X and C(X) is the total number of its outbound links. But page A usually links to other pages itself. Thus, these pages get a PageRank benefit also. If these pages link back to page A, page A will have an even higher PageRank benefit from its additional inbound link. The single effects of additional inbound links shall be illustrated by an example.



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 PageRank of 1. We now add a page X to our example, for which we presume a constant Page Rank PR(X) of 10. Further, page X links to page A by its only outbound link. Setting the damping factor d to 0.5, we get the following

equations for the PageRank 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©=0.5+0.5 PR(B)

PR(D)=0.5+0.5 PR©

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 PageRank values:

PR(A)=19/3=6.33

PR(B)=11/3=3.67

PR©=7/3=2.33

PR(D)=5/3=1.67

We see 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.

3.1 The Influence of the Damping Factor

The degree of PageRank 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 get the following equations for our above example:

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

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

PR©=0.25+0.75PR(B)

PR(D)=0.25+0.75PR©

Solving these equations gives us the following PageRank values:

PR(A)=419/35=11.97

PR(B)=323/35=9.23

PR©=251/35=7.17

PR(D)=197/35=5.63

First of all, we see 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

We remark that the way one handles the dangling node is crucial, since there can be a huge number of them. According to Kamvar et al. [Kamvar et al. 03b], a 2001 sample of the web containing 290 million pages had only 70 million nondangling nodes. This large amount of nodes without out-links includes both pages that do not point to any other page and also pages whose existence is inferred by hyperlinks but not yet reached by the crawler. Besides, a dangling node can represent a pdf, ps, txt, or any other file format gathered by a crawler but with no hyperlinks pointing outside [4]. This initial effect is then propagated even stronger by the links on our site. In this way, the PageRank of page A is almost twice as high at a damping factor of 0.75 than it is at a damping factor of 0.5. At a damping factor of 0.5 the PageRank of page A is almost four times superior to the PageRank 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 PageRank of the page that receives the link and the more evenly distributes PageRank over the other pages of a site.

3.2 The Actual Effect of Additional Inbound Links

At a damping factor of 0.5, the accumulated PageRank of all pages of our site is given by

PR(A)+PR(B)+PR(C)+PR(D)=14

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

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

This time the accumulated PageRank increases by 30. The accumulated PageRank of all pages of a site always increases by

$$(d/(1-d)) \times (PR(X)/C(X))$$

where X is a page additionally linking to one page of the site, PR(X) is its PageRank 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, as, 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 PageRank is propagated to external pages. The justification of the above formula is given by Raph Levien 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, he visits on average (d/(1-d)) pages within that closed system. So, this much more PageRank of the linking page – weighted by the number of its outbound links – is distributed to the closed system. For the actual PageRank calculations at Google, Lawrence Page und Sergey Brin claim 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

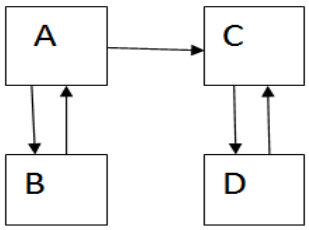
$$(0.85/0.15) \times (PR(X)/C(X)) = 5.67 \times (PR(X)/C(X))$$

So, inbound links have a far larger effect than one may assume. [2]

4. The Effect Of Outbound Links

Since PageRank is based on the linking structure of the whole web, it is inescapable that if the inbound links of a page influence its PageRank, its outbound links do also have some impact.

To illustrate the effects of outbound links, we take a look at a simple example.



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 PageRank of one.[6] Now we add a link which points from page A to page C. At a damping factor of 0.75, we therefore get the following equations for the single pages' PageRank 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 PageRank values for the first site:

$$PR(A) = 14/23$$

$$PR(B) = 11/23$$

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

$$PR(C) = 35/23$$

$$PR(D) = 32/23$$

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

4.1 The Actual Effect of Outbound Links

As it has already been shown, the PageRank 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 PageRank and C(X) is the number of its outbound links. Hence, this value also represents the PageRank 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 PageRank. 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. 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 PageRank of the regarded site diminishes and the pages already receiving a link from that page lose PageRank accordingly.[6]

Even if the actual PageRank 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 PageRank loss of the site, since the above presented formula regards the status after adding the link.

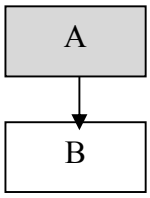
4.2 Intuitive Justification of the Effect of Outbound Links

The intuitive justification for the loss of PageRank by an additional external outbound link according to the Random Surfer Modell 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 PageRank will deplete.

We can conclude that external outbound links diminish the totalized PageRank of a site and probably also the PageRank of each single page of a site. But, since links between web sites are the fundament of PageRank and indispensable for its functioning, there is the possibility that outbound links have positive effects within other parts of Google's ranking criteria. 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.

4.3 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 PageRank cannot be distributed to other pages. Lawrence Page and Sergey Brin characterize links to those pages as dangling links.



The effect of dangling links shall be illustrated by a small example website. We take a look at a site consisting of three pages A, B and C. In our example, 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 get the following equations for the single pages' PageRank values:

```
PR(A)=0.25+0.75PR(B)
PR(B)=0.25+0.375PR(A)
PR(C)=0.25+0.375PR(A)
```

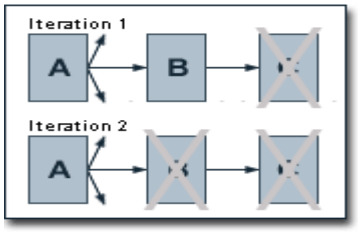
Solving the equations gives us the following PageRank values:

PR(A)=14/23

PR(B)=11/23

PR(C)=11/23

So, the accumulated PageRank 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, the number of dangling links in Google's index is fairly high. A 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 PageRank.



In order to prevent PageRank from the negative effects of dangling links, pages without outbound links have to be removed from the database until the PageRank values are computed. According to Page and Brin, the number of outbound links on pages with dangling links is thereby normalized. As shown in our illustration, removing one page can cause new dangling links and, hence, removing pages has to be an iterative process. After the PageRank calculation is finished, PageRank can be assigned to the formerly removed pages based on the PageRank algorithm. Therefore, as many iterations are needed as for removing the pages. Regarding our illustration, page C could be processed before page B. At that point, page B has no PageRank yet and, so, page C will not receive any either. Then, page B receives PageRank from page A -and during the second iteration, also page C gets its PageRank. Regarding our example website for dangling links, removing page C from the database results in page A and B each having a PageRank of 1. After the calculations, page C is assigned a PageRank of 0.25+0.375PR(A)=0.625. So, the accumulated PageRank does not equal the number of pages, but at least all pages which have outbound links are not harmed from the dangling links problem. The definition of PageRank above has another intuitive basis in random walks on graphs. The simplified version corresponds to the standing probability distribution of a random walk on the graph of the Web. Intuitively, this can be thought of as modeling the behavior of a “random surfer”. The “random surfer” simply keeps clicking on successive links at random. However, if a real Web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will continue in the loop forever. Instead, the surfer will jump to some other page [5]. By removing dangling links from the database, they do not have any negative effects on the PageRank of the rest of the web. Since PDF files are dangling links, links to PDF files do not diminish the PageRank of the linking page or site. So, PDF files can be a good means of search engine optimization for Google.

5. Conclusion

So what we conclude from here is the main factors influencing the page rank is the inbound links and the outbound links including the dangling links. Future work that can be done is the total no of pages affecting the page rank of a web site.

References

[1] The PageRank Citation Ranking: Bringing Order to the Web (PDF, 1999) by Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd.

[2] Sergey Brin, Larry Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Proceedings of the 7th international conference on World Wide Web (WWW)*.
[3] Taher Haveliwala and Sepandar Kamvar. (March 2003). "The Second Eigenvalue of the Google Matrix" *Stanford University Technical Report*: 7056.
[4] Gianna M. Del Corso, Antonio Gullí, Francesco Romani (2005). "Fast PageRank Computation via a Sparse Linear System".
[5] What can you do with a Web in your Pocket (PS, 1998) by Sergey Brin, Rajeev Motwani, Larry Page and Terry Winograd.
[6] Efficient Crawling Through URL rdering (PDF, 1998) by Junghoo Cho, Hector Garcia-Molina and Lawrence Page.
[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The Page- Rank citation ranking: Bringing order to the web. *StanfordDigital Libraries Working Paper*, 1998.

Hitesh KUMAR SHARMA is an Assistant Professor in University of Petroleum & Energy Studies, Dehradun. He has published 8 research papers in National Journals and 5 research papers in International Journal. Currently He is pursuing his Ph.D. in the area of database tuning**.**