Influencing the PageRank using Link Analysis in SEO

Puspita Datta
Research Scholar, Department of Computer Science, Christ University, Hosur Road, Bhavani Nagar, Bengaluru, Karnataka-560029, India. Orcid: 0000-0001-5399-6970

Prof Vaidhehi V
Associate Professor, Department of Computer Science, Christ University, Hosur Road, Bhavani Nagar, Bengaluru, Karnataka-560029, India.

Abstract
The World Wide Web, along with being a boon, can also be a bane since the information contained in it is growing at an alarming rate, and in order to find the right information – the informative content, [6] the search engines have to be highly optimized. There has always been competition between various search engines like Google, Yahoo! and Bing. Although the three work on similar principles, they are very different when it comes to their ranking algorithm. We as users will always use the best search engine which provides the right content in least amount of time and also has a high performance rate and that is only possible if our search engine is optimized. There are certain factors which have great significance in optimizing the search engine. In this paper we are observing the performance of various factors such as – incoming links, outgoing links and broken links and their influence on the ranking of a webpage.

Keywords: Search engine optimization, broken (dead) links, Outgoing links, Incoming links, Adjacency matrix

INTRODUCTION
The opposition in the current world between positioning site pages lie amongst Google and Bing after Bing started powering Yahoo!. But as recent report of April 2017 suggests, according to Net Market Share [1] Bing only serves 7.31% of the global web searches whereas Google serves 77.43% of it. It was found that majority of web-users never thought of researching or finding out something on Internet without using the Google Search Engine. Why is that so? The Google Search has been invented in 1997 and is continually growing with innovative new results, propelled components, and incorporation with other factors.

Let us see an example of how different the search results of similar queries on different search engines like Google and Bing are. If the query string is “best hangout places in Bangalore”, in the homepage of Google which yielded a result of 28,00,000 results in 0.73 seconds whereas when typed in Bing gave 11,10,000 results. It is also noticed that the same query in similar websites were ranked or ordered differently in different search engines, for example www.thrillophilia.com was ranked 6th in Google’s search engine whereas the same website was ranked last in Bing’s search engine. Again www.foursquare.com was ordered 4th rank in both the search engines.

Table 1: The top ten results for search query “best hangout places in Bangalore” at www.google.com and www.bing.com on 5th September, 2017.

<table>
<thead>
<tr>
<th>Order</th>
<th>Google</th>
<th>Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.quora.com">www.quora.com</a></td>
<td><a href="http://www.about.com">www.about.com</a></td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.traveltriangle.com">www.traveltriangle.com</a></td>
<td><a href="http://www.boldsky.com">www.boldsky.com</a></td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.holidify.com">www.holidify.com</a></td>
<td><a href="http://www.blog.travelwithsmile.com">www.blog.travelwithsmile.com</a></td>
</tr>
<tr>
<td>4</td>
<td><a href="http://www.foursquare.com">www.foursquare.com</a> (#)</td>
<td><a href="http://www.foursquare.com">www.foursquare.com</a> (#)</td>
</tr>
<tr>
<td>5</td>
<td><a href="http://www.zomato.com">www.zomato.com</a></td>
<td><a href="http://www.dnaindia.com">www.dnaindia.com</a></td>
</tr>
<tr>
<td>6</td>
<td><a href="http://www.thrillophilia.com">www.thrillophilia.com</a> (*)</td>
<td>m.answers.yahoo.com</td>
</tr>
<tr>
<td>7</td>
<td><a href="http://www.wheelstreet.com">www.wheelstreet.com</a></td>
<td><a href="http://www.romanticbug.com/">www.romanticbug.com/</a></td>
</tr>
<tr>
<td>8</td>
<td><a href="http://www.aapkatimes.com">www.aapkatimes.com</a></td>
<td><a href="http://www.grabbhouse.com">www.grabbhouse.com</a></td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.tripadvisor.in">www.tripadvisor.in</a></td>
<td><a href="http://www.asklaila.com">www.asklaila.com</a></td>
</tr>
<tr>
<td>10</td>
<td><a href="http://www.tourmyindia.com">www.tourmyindia.com</a></td>
<td><a href="http://www.thrillophilia.com">www.thrillophilia.com</a> (*)</td>
</tr>
</tbody>
</table>

Also, an online check to evaluate the performance of the webpages was conducted on https://www.webpagetest.org/ to analyze which webpage will render first when given the same query “best hangout places in Bangalore” in both www.google.com and www.bing.com

Each webpage had three test runs and 7 parameters to find out the difference between them.

For https://www.bing.com/search?
Table 2: The performance results for loading a webpage in www.bing.com on 7th September,2017 is shown below.

<table>
<thead>
<tr>
<th>PERFORMANCE RESULT</th>
<th>Page</th>
<th>Content</th>
<th>First</th>
<th>First</th>
<th>Document</th>
<th>Fully loaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load time</td>
<td>Byte</td>
<td>Start</td>
<td>render</td>
<td>Speed</td>
<td>index</td>
<td>interactive(beta)</td>
</tr>
<tr>
<td>1.205s</td>
<td>0.251s</td>
<td>0.367s</td>
<td>907</td>
<td>3.132s</td>
<td>1.205s</td>
<td>89KB</td>
</tr>
</tbody>
</table>

But for https://www.google.co.in/search

Table 3: The performance results for loading a webpage in www.google.com on 7th September,2017 is shown below.

<table>
<thead>
<tr>
<th>PERFORMANCE RESULT</th>
<th>Page</th>
<th>Content</th>
<th>First</th>
<th>First</th>
<th>Document</th>
<th>Fully loaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load time</td>
<td>Byte</td>
<td>Start</td>
<td>render</td>
<td>Speed</td>
<td>index</td>
<td>interactive(beta)</td>
</tr>
<tr>
<td>2.049s</td>
<td>0.418s</td>
<td>0.567s</td>
<td>1097</td>
<td>&gt;1.744s</td>
<td>2.049s</td>
<td>511KB</td>
</tr>
</tbody>
</table>

In the initial phase Bing had a lesser load time as compared to Google that means initially webpages and websites get loaded quicker in Bing than in Google. But this process slows down when it comes to the intermediate phase. Lastly, Google renders the fully loaded document by using lesser memory, lesser number of page requests and in much less time as compared to the Bing search engine. Thus, Google is a better search engine than Bing in terms of fully loading the document on the net.

The major contributions in this paper are:

- A comparative study between Google and Bing’s search engine and analyzing their performance.
- Finding how incoming and outgoing links influence the PageRank of a webpage
- Proposing an algorithm for broken links and showing their negative influence on the PageRank.
- Consideration of cases that define the factors influencing the overall PageRank.

PROBLEM DEFINITION

Search engines should be optimized for their better performance. The factors affecting their performance rate are huge in number. Some of these factors have a positive or favourable effect on the PageRank of the webpages whereas some have an adverse or negative effect on the ranking of the pages. While linking two webpages, say A and B, two concepts are mainly used – outgoing links and incoming links. They are both advantageous as they add credibility to the webpages. Another concept which is often considered in link analysis is broken links. These links have a negative effect on the ranking of webpages. Thus, this paper aims in demonstrating the influence of these factors on the page ranking of different webpages.

RELATED WORK

Major differences between Google and Bing’s search engine and their performances while ranking webpages:

Google’s head start in 1997

Bing is the name given by Microsoft as its web crawler, already called Windows Live Search and MSN Search. Bing fully started off in the year 2005, intending to provide correct information about more certifiable settings than simply discovering the content on a page whereas Google Search has been found in the year 1997. Bing additionally controls Yahoo’s pursuit and serves only 7.31% [1] of global web searches.

Click through rate

Click through rate or as commonly known as (CTR) in Computer Science is the success ratio which determines the number of clicks made on that particular website.

Now the drawback with Bing’s Search engine is that it is too plain and exact with words. The outcome that is client goal, is frequently missed for the genuine words being written in.

Google, changed the way clients seek its Hummingbird refresh. Google understands that common dialects and profits come from the setting of the words being entered into the hunt bar, and not only the words themselves. Google comprehends and gets equivalent words, and concentrates on singular keywords. [7] The results from a recent study showed that for Google CTR was 18.2% for a No. 1 rank and 10.05% for a No. 2 rank whereas results from the Bing study showed that CTR was 9.66% for a No. 1 rank and 5.51% for a No. 2 rank. [2][5]

Facts about backlinks

The nature and number of backlinks a site receives from different destinations decides Google’s PageRank algorithm. Each “clean” backlink is treated by Google as an “upvoter” in the support of the site. What decides the search rankings of your site is based on the level of authority the sites that link back to your site holds. The higher the authority, the better the search rankings. On the other hand, Bing values quality of the sites compared to the quantity. [17] Even though they also use backlinks from other sites to rank but it is not the sole factor deciding the search ranking. In Bing, the inbound anchor text
needs to exactly or at least partially match your site’s keywords for your backlinks to even count unlike Google. To fight the disadvantages of this like spam backlinks for example in Google, the Google’s Penguin update was developed. [18]

Efficient decision making

A determined issue with Bing is that its bots don’t read through your whole page to locate a pertinent keyword. Clearly around 100 kb towards the start of a site page, is the content that they regularly scan. This stands interestingly with Google bots that look over each and every substance of your page, before settling on a choice about your search ranking. When you present your site to a web crawler for being ordered, you anticipate that your whole sitemap will discover a place in its registries. This happens with Google, yet on account of Bing, odds are, just your homepage will be indexed. [9]

Media Rules on Bing

The stunning picture of the day on the Bing homepage is one of the striking features that styles the page. Bing takes this beautiful feel somewhat further and incorporates it with its search results. Google is fundamentally a content driven web crawler, which uses the Alt tags attached to the images while coding the page to make some sense of those images. Flash cannot be seen or be comprehended by Google bots, this affects the Google’s web architecture by a huge margin.

These are some of the many features which demonstrates the difference between Google and Bing as search engines. Although Google dominates the online search with 89.44% market share as per the first quarter of 2016 and Bing is nowhere in competition to the giant Google, but its market share is increasing rapidly with respect to Google. According to the recent reports by comScore, “to be exact, Bing’s market share rose by 0.2% while Google’s dropped by 0.2%.” [3] probably due to windows 10 integration.

Engine Model: We need to keep one thing in mind that web search tools or search engines don’t rank or list websites, they rank and list webpages. [11] That implies each page of your site can possibly be recorded if you set it up legitimately and each web search tool has its own particular formula for determining and ranking of the sites. In this paper, we have used two input factors namely, incoming and outgoing links of a webpage and how these parameters influence the page rank of a website. The results are analyzed by varying these input parameters.

Maximizing the PageRank score of a web page therefore has become an important component of company marketing strategies where as other search engines rely entirely on web page content to determine ranking of results. Brin and Page realized that web page developers could easily manipulate the order of the results which are searched by placing incorrect information within the webpages. [8]

Research Gap – The major gap identified while researching on link analysis was that the broken or dead links were not considered along with outgoing and incoming links.

Factors in link analysis:

Outgoing or Outbound links

Outgoing or external links are the links which either points to different websites from our website or to our own website in its different pages. External linking to other websites from our own website might be a bad idea as we don’t want our visitors to leave the site and enjoy the content provided in other sites. [12] Thus external linking from our website to its own webpages can be beneficial and will help to increase the ranking of our website. However, external linking from our webpage to other websites in order to locate information related to the visitor’s topic can act as a useful resource as it will result in higher user experience. We can also link our webpage to different educational, governmental and non-profit organizational websites.

Incoming or Inbound Links

Incoming links are the links that are pointing to our website from different websites. The more the links, the better, as this demonstrates to the search engine about how other websites find the content you contain as worthy of offering to their own visitors. You are suggested as an authority on the subject by them. Thus obtaining links for our website becomes very necessary for the ranking of our website. But doing so is not very easy. [4] A link from a highly ranked website has more value than a link from a less ranked website. In other words, well known websites have higher weightage and if they have links which are directed to our webpage then that means our

Figure 1: Outgoing links of Webpage A to Webpage B, C and D
webpage has some quality content in it and so in return it will increase the rank of our webpage among the list of different webpages. Again, the more appropriate the content provided by the link, the more valuable the webpage becomes. [12]

![Diagram of incoming links: A to B, A to C, A to D]

**Figure 2:** Incoming links of Webpage A from Webpage B, C and D

**Broken Links or Dead links**

Broken links or Dead links are the links from a webpage that doesn’t work anymore. This might occur when our website encounters the following situations:

- If an improper URL has been entered or the structure of the URL has been changed.
- If the website is removed or doesn’t exist.
- If the destination webpage from what it is linked to is removed.

In these situations, a 404 HTTP response code (page error) occurs. Broken links to our website or within our websites, both can hamper the ranking of our website [16] and thus it negatively affects to the growth of the website. Hence, updating them continuously and removing some of them becomes an important task as it improves user experience and quality of our website. Visitors would not prefer reading content form a site which shows an error and would search for other sites to receive their content. [15] [16]

**PROPOSED MODEL**

In this segment, we will describe about how the factors of search engine optimization in link analysis influence the PageRank. The complete structure has been discussed below.

**Search Engines:** Before going deep into the working of how Google’s PageRank algorithm works lets first get a brief introduction about what a search engine is and why is it important?

A Search engine is a program or a set of programs which searches and recognizes individual units in a database which is similar or has close similarity with the word or character specified by the user. [4] This is mainly used for locating particular sites on the World Wide Web. The full form of SEO is Search Engine Optimization. It is a continuous process to obtain “organic”, “editorial” or “natural” traffic from the search results through search engines and that is done by optimizing webpages or websites as a whole in order to make them search engine friendly, thus achieving higher ranks in search results. [13]

![Diagram of workflow of search engines]

**Figure 3:** Workflow of Search Engines

**PROPOSED ALGORITHM**

The first webpage ranking calculation known as the PageRank was demonstrated by Lawrence Page and Sergey Brin by counting the number of links (both incoming and outgoing) and the quality of each of those links. PageRank determines which webpage is more important, based on the assumption that more important webpages receive more links than less important webpages. [13]

It is given by

\[
PR(A) = (1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)
\]

where

\[PR(A)\] is the PageRank of the webpage \(A\),

\[PR(T_i)\] is the PageRank of the webpages \(T_i\) which are linked to webpage \(A\),

\[C(T_i)\] is the number of outbound or outgoing links on webpage \(T_i\).

On the initial PageRank algorithm, we add another factor that is the number of broken links from each webpage which affects the page ranking of each webpage negatively and is given by

\[
PR(A) = (1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right) - (b* B(T_i))
\]

On the initial PageRank algorithm, we add another factor that is the number of broken links from each webpage which affects the page ranking of each webpage negatively and is given by

\[
PR(A) = (1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right) - (b* B(T_i))
\]

where

\[PR(A)\] is the PageRank of the webpage \(A\),

\[PR(T_i)\] is the PageRank of the webpages \(T_i\) which are linked to webpage \(A\),

\[C(T_i)\] is the number of outbound or outgoing links on webpage \(T_i\).
B(Ti) is the number of broken links on webpage Ti
d is the damping factor which can be set in the vicinity of 0 and 1 and
b is the broken link factor which is set between 0 and 0.1 (as the ranking of the webpages lie between 0-1)

From the above given equation 1 we understand that the PageRank algorithm does not rank websites all together, rather it ranks each page one at a time. In other words, each webpage is ranked exclusively. Moreover, the PageRank of webpage A is recursively characterized by the page ranks of those pages which are linked to webpage A. [14]

Inside the PageRank calculation, the PageRank of a page T is constantly weighted by the quantity of outbound links C(T) on the webpage T. The weighted PageRank of pages Ti is then included. The result of this is an extra inbound connection for page A which will dependably expand page A's PageRank.

At last, the weighted Page Ranks of all the pages Ti is increased by multiplying with a damping factor d which can be set in the vicinity of 0 and 1. Thus by having too many outbound links, page ranking is lessened.

EVALUATION METRICS AND RESULTS
We now evaluate our proposed method in order to see the effect of each of the factors on the page rank. We have divided the problem into three different cases for better understanding of each situation.

Case 1: When the number of outgoing links of each of the webpages is same, that is when outgoing links of all the webpages is equal to the number of incoming links of all the webpages and there are no or same number of broken links from each webpage then the rank of the different webpages remains the same.

Let us consider 5 webpages A, B, C, D, and E respectively. If the number of incoming links equals the number of outgoing links for all the webpages, then their adjacency matrix will be:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

As discussed above, the rankings of each of the webpage are same. Below figure 4 shows the cluster of back bone nodes and with multiple sink nodes.

![Figure 4: Same ranking of all the webpages](image)

As
PR(A)= PR(B)+PR(C)+PR(D)+PR(E)
PR(B)= PR(A)+PR(C)+PR(D)+PR(E)
PR(C)= PR(A)+PR(B)+PR(D)+PR(E)
PR(D)= PR(A)+PR(B)+PR(C)+PR(E)
PR(E)= PR(A)+PR(B)+PR(C)+PR(D)

So, PR(A)=PR(B)=PR(C)=PR(D)=PR(E)

Case 2: When the number of outgoing links of each of the webpages is same, that is when outgoing links of all the webpages is equal to the number of incoming links of all the webpages and there are broken links then ranks of different webpages differs.

Let us consider 5 webpages A, B, C, D, and E respectively. If the number of incoming links equals the number of outgoing links for all the webpages, then their adjacency matrix will be:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Adjacency matrix of 5 webpages when incoming links equals outgoing links

Table 5: Adjacency matrix of 5 webpages when incoming links equals outgoing links
Figure 5: Shows the ranking of webpages when there are broken links from each webpage

As shown in the figure 5, when we add broken links to the webpages the ranks differ. As we already know that broken links affect the rank negatively so we try to reduce the number of broken links from each webpage. The webpage having a higher number of broken links will have a lower rank. Thus here, in this example since the other linking factors are same so to understand the difference we have given different number of broken links to the different webpages. Starting from webpage 1 which has the least number of broken links, followed by webpage 2 and then so on.

Case 3: When the number of outgoing links of each of the webpages is same, that is when outgoing links of all the webpages is equal to the number of incoming links of all the webpages and there may or may not be broken links then ranks of different webpages differs.

Let us consider 3 webpages A, B and C respectively. As shown in the adjacency matrix below webpage A has 2 incoming links and 1 outgoing link, webpage B has 1 incoming link and 2 outgoing link and webpage C has 1 incoming link and 1 outgoing link.

**Table 6:** Adjacency matrix of 3 webpages

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

From here we can conclude that webpage C with 1 incoming and 1 outgoing link has the best rank among the 3 webpages while webpage A is ranked second as it has 2 incoming links and 1 outgoing link, as more the number of incoming links more is the PageRank. Also, for webpage C which has 1 incoming link and 2 outgoing links so its ranked the last as more number of outgoing links affect the PageRank score of the webpage.

CONCLUSION

PageRank is a very critical segment of Google’s indexing of pages, but it is needless to mention that it is not the only segment. While observing the first example it is noticed that www.tripadvisor.in is the ninth result in Google’s top ten results. There are six results which are considered more pertinent by Google to the given query “best hangout places in Bangalore” that have cut down toolbar PageRank scores than www.tripadvisor.in.

As a high PageRank score for a page doesn’t ensure that the website page appears high in the positioning of list items, so website design and numerous different factors are imperative for the improvement of the website's positioning, for instance, placing and repetition of relevant words to the visitor's topic must be considered while making the pages. The reality of the situation is most web index clients want to find appropriate information quickly at any point of time. To ensure that the clients are satisfied, Google must keep into consideration that the most substantial site page shows up at the highest point of postings. To stay focused, organizations and media must find out an informative way to reach till there.

REFERENCES

[5] Nicole Immorlica, Kamal Jain, Mohammad Mahdian, and Kunal Talwar, “Click Fraud Resistant Methods for Learning Click-Through Rates”, Microsoft Research, Redmond, WA, USA
Department of Computer Science Ithaca, NY 14853 USA


[16] https://www.computerhope.com/jargon/b/broken_link.htm


[18] https://searchengineland.com/library/google/google-penguin-update