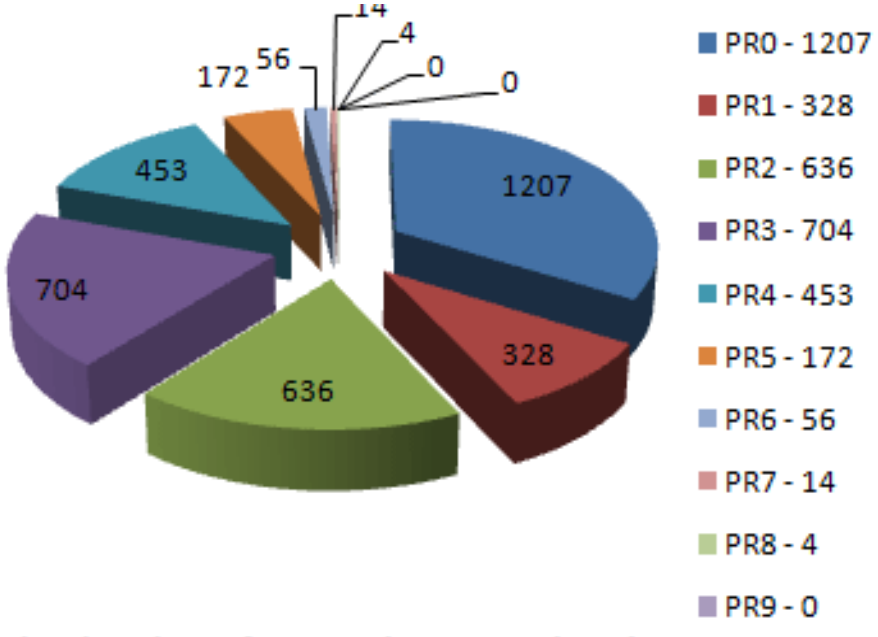
Google Page Ranking Algorithm

Shubham Upadhyay (P2010CS1036)



Directories situations after Google Page Rank update

Introduction

Within past few years, Google has become the far more utilized search engine worldwide. A major factor was the superior quality of search results which is based on the Google's page ranking algorithm.
Page Rank algorithm is the link analysis based algorithms named after Larry Page & Sergey Brin and used by the Google Internet search engine, which assigns a numerical weighting to the web pages. The name "Page Rank" is a trademark of Google. Stanford University has the patent to it.
Page Ranking is an algorithm which gives a numerical value to each page on the web. It does not rank websites as a whole rather it ranks individual pages. The importance or the ranking of a page improves when a page is found with a link to this page. Hence with every page with one outbound link to this page piling on, the rank of this page improves. It improves further if a page with higher rank has a link to this page.
Using this Page Ranking algorithm, Google is providing its one of the best search results showing higher rank pages containing the keyword searched as the top choices followed by the lower rank pages.

Page Ranking

Page Ranking Algorithm is based on the Random Surfer Model which is based on one random surfer who is surfing the web by simply clicking on the links with no regard towards content. Certain probability is associated with the event of surfer visiting a page. This probability is also derived from the page's page rank. The Probability that the surfer clicks on one link is solely given by the number of links on that page.
Random Surfer could click on one of the links on that page or he could stop clicking on a link on that page due to boredom or any other cause and select another page at random, this probability of him doing so, is also known as the damping factor.
Hence the formula for calculating the page rank is proportional to the sum of fraction of page rank of all the pages which has an outbound link to that page with respect to the number of outgoing links multiplied by the probability of him choosing a link on that page or damping factor. Larry Page and Sergey Brin in their research paper published that on an average, the damping factor comes out to be 0.85. Page Rank is also improved when a page has a lot of inbound links from pages with higher page ranks.

Facts

- If popular sites have links to a particular page, then the chances of that page having a higher page rank increases.
- A link from the Google's trusted web pages like .gov or .edu would lead to increased page rank.
- Link from the similar content web pages would increase page rank i.e. if your friend doesn't vote for why should others?
- Sometimes guilt from association can be very damaging to your social status. Google punishes web pages which have a link from spam sites like link farm.
- Mutual linking between web pages cancels their effectiveness in page ranking.

An Analogy

Page ranking algorithm can also be used as Social Ranking Algorithm on a social networking site. As we are ranking web pages, one can also rank people's profile depending on the number of friends they have. Social Rank not only depends upon the number of friends rather it would depend upon the probability of one clicking on one's profile from the profile of theirs friends.
A Friend with high rank will increase this person's rank. Due to the similarity with page ranking algorithm, this technique had been effectively used by the sites like Facebook to give ranking to the people's profile. As one searches for a particular keyword on Facebook, the search results are in decreasing order of rank.

Formula

$$\mathbf{PR(A) = (1 - d) + d * \sum_{i} \frac{PR(T_i)}{L(T_i)}}$$

- ***d*** → Damping Factor (usually 0.85)
- ***PR(A)*** → Page Rank of page A
- ***Ti*** → i^{th} page containing the link to the page A
- ***L(T_i)*** → Total number of outbound links on page Ti

Interesting part of Algorithm

The most interesting part of Page Ranking is the introduction of PR0 or Zero page rank. The pages with PR0 are not thrown out of index but rather they are at the end of the search results. The causes behind a page having PR0 are mainly not having many inbound links from the pages with higher ranks.
Google has used this concept of PR0 to detect Spam.
Spam has always been the biggest problem that the search engines had to deal with. When a Spam is detected, the usual proceeding is the banishment of the websites which is the difficult task on its own. If a page has many outbound links to the less rank pages then its rank will surely decrease. Many bad neighbors could lead to PR0 with a maximum chance to be termed as a Spam.
Google punishes web pages which have a link from spam sites like link farm.

Significance

World Wide Web has been responsible for providing us with all the useful information which many of us take for granted. Presently there exist nearly 150 million websites on the web space. Finding useful information after searching through this huge stack of websites is an enormous task to perform. That's why most of the search engines use complex algorithms that search all the websites for the useful information.
With the use of this algorithm, Google's been searching websites containing the keywords which users search and based on it, ranking the web pages. With this, Google has been efficiently providing its excellent search results with higher rank pages on the top and the pages with lower rank following.
This has enormously decreased the searching time of the useful information on the web space. With the use of this algorithm in the spam filtering processes of web sites has given a protective environment to the users so that the users could not be misled towards bad websites or wrong information.

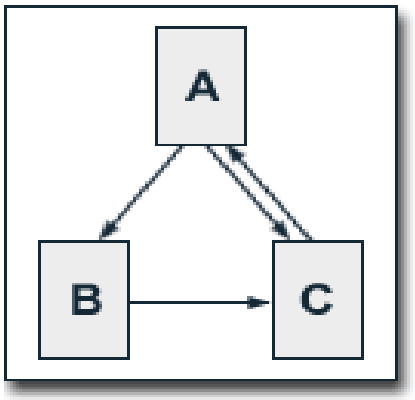
Additional Factors Influencing Rank

Larry Page himself pointed out later that there can be some more factors which can influence the page rank. The factors influencing page rank are:

- Visibility of a link
- Position of the link on the page
- Distance between web pages
- Importance of a linking page

These factors can have certain effect on the page's Rank like less visibility of the link decreases the probability of the random surfer making the choice of the link to that page thereby decreasing the page rank of that page. Position of the link also marks the visibility of that link. Larger distance between the web pages decreases the probability of that link to be picked. Importance of a linking page is one of the major factors in deciding the page's rank as a link from the more important page increases the page rank of the page of which link is present on this page.

Example



We have a small example to illustrate the algorithm consisting of three pages A, B and C whereby page A links to B and C, page B links to page C and page C links to page A. To simplify calculation, let damping factor be 0.5. According to the equation,

$$PR(A) = 0.5 + 0.5 * PR(C)$$

$$PR(B) = 0.5 + 0.5 * \left(\frac{PR(A)}{2}\right)$$

$$PR(C) = 0.5 + 0.5 * \left[\left(\frac{PR(A)}{2} + PR(B)\right)\right]$$

Solving these equations, we get
PR(A)=14/13=1.07692308
PR(B)=10/13=0.76923077
PR(C)=15/13=1.15384615

The sum of the page ranks is equal to the number of web pages, in this case 3.

Implementation

In practice, calculation of the solution to these equations is a difficult task even on a computer. So there is even a simple way to get the solution which is also known as the "Iterative approach". Due to large number of pages on the web, this is an approximated approach which assigns a starting value to all the pages and calculates the corresponding page rank values using these equations iteratively for 100 iterations. According to publications of Larry Page and Sergey Brin, for a good approximation of the page rank values of all the web pages on the internet, about 100 iterations are necessary. Implementation of this approach for the above mentioned three page example is as follows:

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

As we can see in 12 iterations, we have got the same solution as above.

Conclusion

Google's page ranking algorithm has benefited all of the users with its quick and efficient search results. This algorithm has saved the most important time of the users and protecting the users from mal sites. Algorithm has an adaptive nature and has been modified for inculcating several additional factors which influence page's page rank.

Collaborator

Nitish Duggal P2010CS1029

Bibliography

- en.wikipedia.org/wiki/PageRank
- google.com/technology/pigeonrank.html
- computer.howstuffworks.com/google-algorithm.htm
- pr.efactory.de/e-pagerank-algorithm.shtml
- http://infolab.stanford.edu/~backrub/google.html