PageRank Algorithm Overview

Ilija T Jolevski[1], Gjorgi K. Mikarovski[2] and Aleksandar N. Kotevski[3]

***Abstract* – This article introduces and discusses the concept PageRank link analysis algorithm, brief history, which factors do have an impact on PageRank, Based on research to several web portal, guidelines are provided on how to optimize web page context for search engines. In addition, we discuss for bad practice for over-optimizing articles posted to web page.**

***Keywords* – PageRank, optimize, SEO, algorithm**

I. INSTRUCTIONS FOR THE AUTHORS

Authors have two main goals: articles to be available for search engines users and where articles are displayed in the results list – articles which are displayed in top positions are more likely to be read. They should have an interest in ensuring that their articles are indexed by search engines like google, yahoo, ask, msn because they have two main goals: articles to be available for search engines users and where articles are displayed in the results list – articles which are displayed in top positions are more likely to be read. The most search engines use crawlers to find pages for their algorithmic search results. Pages that are linked from other search engine indexed pages do not need to be submitted - they are found automatically. Usually used link analysis algorithm is pagerank – used by google search engine. Exploiting the hyperlink structure of the web, pagerank surmises that each web page has a prestige score that ties to the number of in-links the page receives. It work based on assigns a numerical weights to each member of set of hyperlinked elements of documents. Assigned numerical weight that is assign to element e is also known as pagerang of e and denoted by pr(e). Pagerank can be calculated for any collections of documents of any size. With pagerang view, some page a has a higer pagerang than another page b, even though it has fewer links to it – the link it has is of a much higer value. A hyperlink to a page counts pagerang detect as a vote of support. So, if some page is linked to many pages with high pagerang, that it received a high rang itself. Contrary, of there are no links to a web page there is no support for that page.

[1]Ilija T Jolevski is with the Faculty of Technical Sciences, I.L.Ribar bb, 7000 Bitola, Macedonia, E-mail: ilija.jolevski@uklo.edu.mk

[2]Gorgi K. Mikarovski is with the Faculty of Technical Sciences, I.L.Ribar bb, 7000 Bitola, Macedonia, E-mail: gjorgi.mikarovski@uklo.edu.mk

[3]Aleksandar N. Kotevski is with the Faculty of Low Sciences, Prilepska bb, 7000 Bitola, Macedonia, E-mail: aleksandar.kotevski@uklo.edu.mk

II. HISTORY OF PAGERANG ALGORITHM

PageRank was invented at Standford University in 1995 year, by Larry Page and Sergey Brin, as a part of research project for developing a new kind of search engine. The first publication about this project, describing PageRank and the initial prototype of the Google search engine, was introduced in 1998 year. Shortly after, Google Inc. was founded to manage the Google search engine, which used PageRank. Currently the trademark PageRank belongs to Google Inc., but the original patent for the PageRank algorithm is assigned to Stanford University.

The PageRank algorithm is still one of the factors taken into account in Google's search engine result ranking, and it is constantly followed by interested search engine marketing specialists around the world.

III. WHAT IS PAGERANG ALGORITHM

PageRank is algorithm formulated by Sergey Brin and Larry Page, and today it is used by Google search engine. According to PageRank algorithm, importance of any web page can be define as number of links pointing to it from other web pages. Google uses PageRank to adjust results so that sites that are deemed more "important" will move up in the results page of a user' s search accordingly.

PageRank can be understand as number assessed solely the voting ability of all incoming links to a page, and also, how much they recommend that page. Every unique page of every site that is indexed in Google has own PageRank and it is one of the most important ranking techniques used in today's search engines. Not only is PageRank a simple, robust and reliable way to measure the importance of web pages but it is also computationally advantageous with respect to other ranking techniques in that it is query independent, and also content independent.

IV. PAGERANG CALCULATION

The basic PageRank's idea is if one web page A has a link to another web page B, then the author of A is implicitly conferring some importance to web page B. The first model for PageRank uses structure of links of the web pages to construct Markov chain with primitive transition probability matrix M. Irreducibility on chain is guarantees for existing of PageRank vector. If there is one set from x web pages and one matrix M. PageRank algorithm as first constructs new matrix N by renormalizing each row of matrix M to sum 1. If there is some web surfer who visit some web page, and decides for browsing to next web page. There are two possibilities: user

should click on hyperlink that is on current page, with probability 1-e, or, with probability e, user should enter new URL address at browser. In this case, e is parameter, set to 0.1-0.2. This process is Markov chain on the web pages, with transition matrix eU + (1-e)N. In formula, U is transitions matrix. R, the scores on vector of PageRank must to be defined to be the stationary distribution of this Markov chain. Also, R can be define as

$(eU + (1 – e)M)^T r = r$

After this, authoritativeness of page i is the asymptotic chance of visiting page.

Actually, PageRang algorithm is elegantly simple, and is calculation is following:

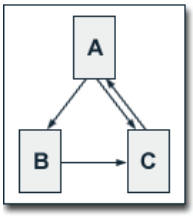
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

In this calculation, PR(A) is the PageRank of a page A, PR(T1) is the PageRank of a page T1, C(T1) is the number of outgoing links from the page T1, d is a damping factor in the range 0 < d < 1. D is usually set to 0.85.

It is natural to wonder what is the best value of the damping factor, if such a thing exists. In a way, when ® gets close to 1 the Markov process is closer to the "ideal" one, which would somehow suggest that ® should be chosen as close to 1 as possible. This observation is not new, but it has some naivety in it.

There are two ways in which this algorithm can affect to position of page on Google. The number of incoming links is important for web rating – the more of these links has positive effect. If some page hasn't incoming links, it can have a negative effect of web ratings. Contrary to incoming links, more outgoing links has negative effect of page rang.

Of there is some web consisting of three pages, A,B and C. Here, page A links to the pages B and C,page B links only C, and page C links only page A.



For this example, we have following calculation if we assume that d factor is 0.5:

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

PR(B) = 0.5 + 0.5 (PR(A) / 2)

PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following PageRank values for the single pages:

PR(A) = 14/13 = 1.07692308

PR(B) = 10/13 = 0.76923077

PR(C) = 15/13 = 1.15384615

From calculation, is obvious that sum of PageRanks of all pages is 3- that is equals the total number of web pages. Here is easy for calculation because of very small structure of web page. But, today web consist of a huge number of documents, and practically

It is obvious that the sum of all pages' PageRanks is 3 and thus equals the total number of web pages. As shown above this is not a specific result for our simple example.

For our simple three-page example it is easy to solve the according equation system to determine PageRank values. In practice, the web consists of billions of documents and it is not possible to find a solution by inspection. Therefore, Google search engine uses iterative computation of PageRank values. Iteratuve computation of PageRank means that each page is assigned an initial starting value. PageRanks of all pages are then calculated in several computation circles based on the equations determined by the PageRank algorithm. For the same example, the iterative calculation look like:

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

From computation above, we can conclude that real PageRank values can be compute in only few iterations. Appropriation, and according of Lawrence Page and Sergey Brin, nearly 100 interactions will compute a good approximation of the Page Rang values of the whole web.

How is PageRang determinate?

In situation from 2 web page, A and B, if A links to page B, then Page A is saying

that Page B is an important page. Also, if a page has important links pointing to it, then its links

to other pages also become important, irrespective of the actual text of the link.

Simplified algorithm

In situation with four web pages, A,B,C and D. Each of these pages would begin with an estimated PageRank of ¼ (0.25).

If pages B,C and D link to A, then they would each confer 0.25 PageRank to A. All PageRank PR( ) in this simplistic system would thus gather to A because all links would be pointing to A.

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

This is 0.75.

If B also has a link to page C, and page D has links to A,B and C. The *value of the link-votes is divided among all the outbound links on a page*. Thus, page B gives a vote worth 0.125 to page A and a vote worth 0.125 to page C. Only one third of D's PageRank is counted for A's PageRank (approximately 0.083).

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

PageRank can be calculated as document's own PageRank score divided by the normalized number of outbound links L( ).

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

In the general case, the PageRank value for any page u can be expressed as:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}.$$

An important aspect of PageRank calculation is the matrix inversion. Therefore, we'll discuss about few numerical inversion algorithms:

- Jacobi iteration

German mathematician Carl Gustav Jacob Jacobi has been developed simple method which is defined by

$x\{0\} = b$
$x\{i+1\} = (1 – M) * x\{i\} + b$

for $i \geq 0$, where {i} denotes the result after the i-th iteration.

Jacobi matrix is the Matrix (1 - M). In case of PageRank calculation the Jacobi matrix is given by d T (damping factor times transition matrix), a sparse matrix. The solution of the iteration is x, if the limit exists. The convergence is guaranteed, if the absolute value of the largest eigen value of (1 - M) is less than one. In case of PageRank calculation this is fulfilled for 0< d < 1.

There is a generalization of the algorithm. Using the decomposition (D - M), where D is an easy to invert matrix (e.g. the diagonal elements of M), leads to:

$x\{0\} = D^{-1} * b$
$x\{i+1\} = D^{-1} * ((D – M) * x\{i\} + b)$

Gauss-Seidel method

An improved algorithm is the Gauss-Seidel iteration. It based on the decomposition

$M = D + L + U$

where D, L and U are the diagonal, lower triangular and upper triangular parts of M. This yields

$x\{i+1\} = D^{-1} * (L * x\{i+1\} + U * x\{i\} + b)$

Introducing a relaxation parameter $\lambda \neq 0$ leads to a generalization of the Gauss-Seidel method:

$x\{i+1\} = (1 - \lambda) x\{i\} +$
$\lambda D^{-1} * (L * x\{i+1\} + U * x\{i\} + b)$

Minimal residue

Another iteration scheme is the minimal residue iteration. It is given by

$x\{i+1\} = x\{i\} + r\{i\} (r\{i\}_j M_{jk} r\{i\}_k) /$
$(M_{jk} r\{i\}_k M_{jl} r\{i\}_l)$

where the residue is defined by

$r\{i\} = b - M * x\{i\}$

The minimal residue iteration is never divergent.

In situation when some random surfer that starts from a random page, and at every time chooses the next page by clicking on one of the links in the current page. So, we could define the rank of a page as the fraction of time that the surfer spent on that page on the average. Clearly, important will be visited more often, which justifies the definition.

PageRank also can be define as the stationary distribution of a stochastic process whose states are the nodes of the web graph. The process itself is obtained by combining the normalised adjacency matrix of the web graph with a trivial uniform process that is needed to make the combination irreducible and aperiodic, so that the stationary distribution is well defined. The combination depends on a damping factor, which will play a major role in this paper. When damping factor is 0, the web-graph part of the process is annihilated, resulting in the trivial uniform process. As damping factor goes to 1, the web part becomes more and more important.
PageRank also can be defined as sum of the PageRanks of all incoming links, divided by the number of its outgoing links.

PageRank is Google's method of ranking individual web pages. Google looks at the pages which link to your page and how they rank in terms of importance. In a nutshell, pages that have links from important, high quality pages, receive a higher PageRank. Google combines PageRank with sophisticated text-matching techniques to find pages which are both important and relevant to your search

The PageRank system is a numerical grade from 0-10. There have been great articles written in the past on how to estimate and compute PageRank, and they can be confusing to some

V. FACTORS THAT IMPACT PAGERANK

- Valid Code – HTML/XHTML for the site should be valid
- Tags - using proper meta tags
- Sitemap – for sites with lots of folders and subfolders, the site should have sitemap with site tree structures linked
- CSS content
- Internal links and external links on the same page may not be splitting the Google PageRank vote equally

- Depending on the location of the link, Google PageRank may be weighted differently
- Multiple links to the same URL from the same page may not each get the same piece of the Page Rank vote
- "Run-of-site" external links, like Blogrolls, may now have diminished PageRank
- Links between domains that Google sees as "related" may have their PageRank significantly damped down. Possibly the same goes for sites that link to sub-domains
- Incoming Links from popular sites are important - If pages linking to you have a high PageRank then your page gains some part of their reputation
- Site can be banned if it links to banned sites – developer must be careful of any out-going links from their site because Google will penalize you for bad links
- Illegal activities like deceptive redirects, cloaking, automated link exchanges, or anything else against Google's quality guidelines will penalize your PageRank and possibly ban your site from Google
- Different pages from a site can have different Page Rank - Search engines crawl and index web pages not websites - therefore your page rank may vary from page to page within your website
- Content is not taken into account when PageRank is calculated - Content is taken into account just when you actually perform a search for specific search terms
- Bad incoming links don't have impact on Page Rank – from where the links come it doesn't matter. Sites are not penalized because of where the links come from

VI. CONCLUSION

As the web grows in size and value, search engines play an increasingly critical role, allowing users to find information of interest. Search engines used huge number of algorithm for ranking the sites. PageRank is a link analysis algorithm used by the Google Internet search engine.

VII. REFERENCES

[1] A. Arasu, J. Novak, A. Tomkins and J. Tomlin, "PageRank Computation and the Structure of the Web: Experiments and Algorithms", Technical Report, IBM Almaden Research Center, Nov. 2001

[2] Taher Haveliwala and Sepandar Kamvar. The condition number of the PageRank problem. Technical Report, Stanford University Technical Report, June 2003

[3] Taher Haveliwala. Efficient computation of PageRank. Technical report, Stanford University Technical Report, October 1999

[4] Chris Pan-Chi Lee, Gene H. Golub, and Stefanos A. Zenios - A fast two-stage algorithm for computing PageRank and its extensions. Technical report, Stanford University Technical Report, 2004

[5] Luca Pretto. A theoretical analysis of google's PageRank. In Proceedings of the Ninth Symposium on String Processing and Information Retrieval, 2002

[6] G. Jeh and J. Widom. Scaling personalized web search. In Proceedings of the Twelfth International World Wide Web Conference, 2003

[7] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin. PageRank computation and the structure of the Web: experiments and algorithms. In Proceedings of the Eleventh International World Wide Web Conference, Poster Track, 2002.