A Survey on Search Engine Optimization using Page Ranking Algorithms

M. Sajitha Parveen[1] T. Nandhini[2] B.Kalpana[3]
[1,2]M.Phil. Research Scholar, Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore.
[3]Professor, Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore
Email:aeroscutemillenium@gmail.com,nandhinimphil2015@gmail.com,kalpanacsekar@gmail.com

Abstract - The survey report titled “Search Engine Optimization using Page Ranking Algorithms” presents various page ranking algorithms for optimizing the search engine results. There are various page ranking algorithms which aids the search engines in listing the pages with higher degree of relevance. The page ranking algorithms discussed in this report are page rank algorithm, HITS algorithm and semantic similarity algorithm.

Keywords *:* Search Engine Optimization, Page rank algorithm, HITS algorithm, Semantic Similarity algorithm.

I. INTRODUCTION

Internet plays a predominant usage role in the modern computing era. There arises a necessity of giving the best results to the internet users. The search engine retrieves the contents and lists them by ranking the web pages. The pages with the highest rank are displayed in the first page followed by the pages with lower ranks in the consecutive pages. There are various page ranking algorithms proposed till date to improvising the search engine results. The first all ever page ranking algorithm proposed by Sergey Brin and Larry Page was Page Rank algorithm. Page rank ranks a page as per the number of links attached to it. But Page rank has the limitation that because of dividing the rank equally among the pages results in low rank values which leads to rank pages with less relevant content to be listed on top of the webpages. Second is HITS *Hypertext Induced Topic Search (HITS)* or *link analysis algorithm*. The basic version of HITS algorithm calculates page rank based on Hubs and authorities.

Later it has been improved by the factors such as fetching the pages using keywords to improve relevancy of the content, number of visits of a specific link, bold as well as italic as parameters to filter out more relevant web pages. There are many versions of HITS algorithm. They are I-HITS(*Improved HITS*), M-HITS (*Modified HITS*), CLEVER algorithm. These algorithms enhances the parameters used in fetching the pages and vary in the method of ranking the web pages. Another type of algorithm is the semantic similarity algorithm in which the meaning of the contents in the web pages are analysed. If the meanings of the contents are more alike, then those pages are taken into consideration for higher rankings.

II. LITERATURE SURVEY

In [1], the author gives an approach of optimizing the search engine results by ranking web documents. It also checks semantic similarity which is necessary for the relevant content retrieval which means that the contents given are more relevant that is required by the user. Algorithms such as HITS and Semantic Similarity have been implemented to compare the web ranking. These algorithms prove to be better in ranking efficiency and improves relevancy of content in results. In [2], the author gives a comparative study about different page ranking algorithms such as page rank and trust rank. The future works include improving the accuracy in bringing the relevancy by requiring a match between the query and the text on the expert page which qualifies the hyperlink being considered. This ensures that hyperlinks being considered are on the query topic. For further accuracy, the result of the steps described above is to generate a listing of pages that are highly relevant to the user's query and of high quality. Further, the features of trust rank algorithm and page rank can be combined. Parveen Rani et. al. focuses on optimizing Search Engine results by framing a new page ranking algorithm. The version of HITS algorithm has been modified and M-HITS (Modified HITS) version is developed. The authors have implemented M-HITS version by introducing six parameters to evaluate page rank. The future works can be done by introducing Artificial Intelligence techniques as a part of search engine optimization.

III. CONCEPT OF SEO

Basically, people are in need of internet to gather multi-faceted information. So, right from the advent of the internet era there are millions and millions of websites available variedly designed for satisfying the needs of various types of users. The network traffic is an impact of the numerous upcoming websites. Due to network traffic there arises a difficulty in visibility of websites for fetching the relevant pages by the search engines. So, the term search engine optimization was coined.

Search Engine Optimization is the process of enhancing the sites’ **visibility** by minimizing the web traffic in search engine’s results. SEO aids search engine to place the contents with **high page ranking** at the **top** of search engine’s results page responding to a user query. So, search engines have to be optimized for efficient retrieval of results to the users. Since the best contents are ranked by the search engine. The search engine produces the results such that the websites links’ consists of the keywords that match with that of the keywords in users’ queries. There are chances that more than one websites’ links contain the same keywords but the content or the information may differ. The aspect of semantic similarity plays a vital role in capturing the content with the highest degree of relevance.

IV SEO TECHNIQUES

There are several techniques for optimizing search engine. Page ranking optimizes the search engine's results hence enhancing the visibility of the sites to the search engine. There are several page ranking algorithms in which the page ranking is calculated on the basis of analyzing many criteria such as link structure, links (inbound as well as outbound links), keywords and content. Those algorithms are listed below. They are

A. Page Rank Algorithm
B. HITS Algorithm
C. Semantic Similarity Algorithm

A. Page Rank Algorithm

Page rank algorithm, was formerly developed by Larry Page and Sergey Brin of Stanford University. Page rank algorithm ranks the pages based on link structure of the web. It divides the rank equally among the pages. But, this method has some limitations such as very low ranks were assigned to the pages. Sometimes, the ranking may go down to zero which may affect the consideration of the page for listing that may contain relevant contents. To overcome the limitations found in the page rank algorithm a variation of the page rank algorithm was found which is known as **"Weighted Page Rank"** algorithm was developed which ranks the links according to the popularity of the links. Popularity of the links or pages means the number of times the pages have been visited. Based on this factor the popularity of the page is determined. The Page rank algorithm consists of two techniques namely simplified page rank and complex page rank which are discussed below. The complex page rank consists of random surfer model.

i. Simplified Page Rank

A simplified page rank gives page rank to a page on voting basis. If a link exists from page A to page B, it is considered as a vote, by page A, for page B. In other words, if a hyperlink exists for A which is an inbound link to B then A votes for B. The search engine looks for more number of votes or links a page receives. It also analyzes the page that casts the vote. These votes give weight to the pages and make other pages also important. The illustration below presents an example of a simplified page rank.

Example:

Consider 4 web pages namely A, B, C and D. If the pages B, C and D link to A, then the page ranks of B, C and D are summed.

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

If a page has a link to the same link, for instance, page B also has a link to page C, and page D has links to all three pages, B cannot vote twice. For that, the page rank of B is given half a vote. Similarly, the page rank of D is given as one third for A.

PR (A) =(PR(B)/2) +(PR(C)/1)+(PR(D)/3)

In generic terms, the above calculation can be represented as follows:

PR(A)=(PR(B)/L(B))+(PR(C)/L(C))+ (PR(D)/L(D))

In the above formula, the Page ranks of the respective pages are divided by the total number of links coming from that page. Since, for the reason no page can have a page rank of 0 the above formula is multiplied by a variable called "*damping factor*" and it produces a value of 1-q to avoid the page ranking value does not equal 0.
So, the page rank formula is given as

PR(A)=(1-d)[(PR(B)/L(B))+(PR(C)/L(C))+(PR(D)/L(D)…..)]d

The above process is iterated until all page ranks stabilize at some point. The search engine takes this page rank for consideration at this stage.

ii. Complex page ranking algorithm

There are chances that the user skips to a random page apart from navigating through the routine link structure. So, in order to notice the random clicks of the user another complex algorithm namely random surfer model has been introduced. This algorithm assumes a probability value based on the number of links it is connected to. Here one page's page rank is divided by the number of links on the page. To maintain a minimum page rank the damping factor is represented as (1-d). The page rank is given by

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

N- Total number of all pages on the web.
d-damping factor

The diagram below depicts an example of web structure with ranking.

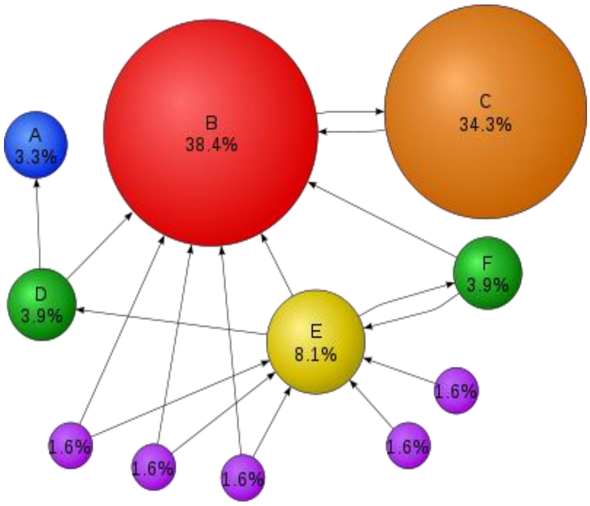

Figure 1 Web Structure with Ranking

Advantages and Disadvantages

Page rank ranks the page with more number of links with higher ranks. So, the pages with more links are considered to be authoritative sources of information. The above factor shows that this is a query independent algorithm. The main disadvantage of page rank algorithm is not a keyword based algorithm because it calculates score only on the basis of hyperlink structure which may tend to lose importance for relevant pages. It is too expensive and time consuming as it calculates scores at after the search process.

B. HITS algorithm

HITS is an acronym of Hyperlink-Induced Topic Search which is also known as link analysis algorithm. It is developed by Jon Kleinberg. This algorithm uses the concept of Hubs and Authorities to calculate page ranks.

Hubs and Authorities

Hubs are links that points to the source of information. Authorities are the source of information. A good hub is one which consists of considerable amount of authorities and a good authority is one which is pointed by a considerable number of hubs. HITS is query dependent. It can be considered as query dependent page ranking algorithm because if a user gives the query as “best car maker” which will retrieve the contents from authority pages such as cars.com or hyundai.com. The pages that contain the links of these websites are hub pages.

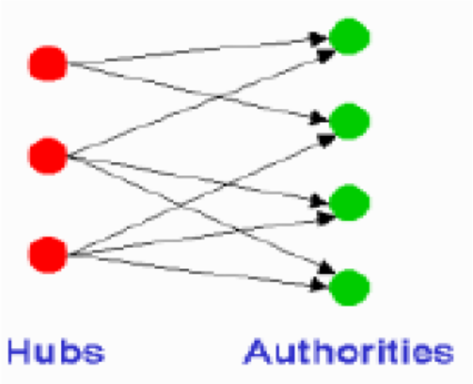

Figure 2 Hubs and Authorities

It works as follows: Whenever the user inputs the query, the search returns a set of web pages with the help of a traditional search engine. In other words, the most relevant pages are retrieved with the help of a text-based search algorithm. This set of pages is termed as *root set*. Again this root set is expanded to give another set of pages known as base set. These are the links which are connected to the root set web pages in which the hyperlinks form a subgraph.

The HITS algorithm computation is based on the focused subgraph. This base set construction ensures that the strongest authorities are included in the focused subgraph.

Working of HITS

First the root node is obtained. Then it is expanded by adding all the pages that are linked to and from the root set forming base set. Then adjacency matrix from base set is computed and iterative eigen vector in the adjacency matrix is computed. Then the top establishment and hubs are reported.

The hub and authority value of page p is calculated using the following formula

Formula 1

$$A_{p=} \sum_{l \in bp} H$$

where A_p represents Authority value of page p. b_p is the set of pages which point to *p* and are present in the base set, and *l* for the number of links.

Formula 2

$$H_{p=} \sum_{l \epsilon b1_{p1}} A$$

where Hp represents hub value of page *p*, $b1_{p1}$ is the set of pages that *p* points to and which are present in the base set and *l* for the number of links.

Then by n adjacency matrix and its transpose AmT is computed. The variables for the factors such as base set which comes as a result of user query as $Gr_{uq,}$ authority values for nodes in Gr_{uq} as Ve_a and hub values as Ve_h are assigned. Then following matrix computation vector is being calculated as follows:

Ve_a = $AmTVe_h$ for formula 1 Ve_h = $AmVe_a$ for formula 2 Substituting values of Ve_a and Ve_h, Ve_a = $AmTAmVe_a$ Ve_h = $AmAmTVe_h$ Thus, Ve_a and Ve_h are the converged principal eigen vectors of AmTAm and AmAmT.

After determining the elements with the largest values in the normalized eigen vector, the top authorities and hubs appear as outcomes. Finally, the authority value is the sum of the scaled hub values that point to that page and the hub value is the sum of the scaled authority values of the pages it points to.

Advantages and Disadvantages

The advantages of HITS algorithm is the popularity of the web pages is revealed. Due to query dependent nature of HITS the content relevance is improved rather than the existing page ranking algorithm. It involves the computation of adjacency matrix with two vectors namely hub and authority. Therefore it is considered as the simplest form in calculation. The concept of hub and authority leads to assignment of high rank to not so relevant pages and topic drift are a major disadvantage of HITS algorithm.

C. Semantic Similarity Algorithm

Semantic similarity algorithm works such that pages with same meaning are taken into account based on keywords which are fetched from different web pages. The algorithm works as follows: “First, the relevant texts are collected and a text-list is constructed on link basis. Then, the user query is made into a single string. Each text in the text-list is created text vector space and domain-dictionary of words using statistical-model () and domain-dictionary and relevance-value of text corresponding to user query is computed. Compute domain-ontology and domain similarity of text value with domain ontology. Now, verify the maximum of the domain-similarity value and the relevance-value. The obtained maximum value is called the relevance-score value. Iterate the above steps starting from creating text-vector space up to computing relevance score till no more text is left. Organize the text according decreasing order of relevance score and assign rank to them. Finally, display the contents according to their ranks” Nisha et.al.

Advantages and Disadvantages

Semantic Similarity algorithm proves to be a text based ranking algorithm along with the meaning of the pages are considered. The disadvantage is due to string consideration there exists the problem of relevancy mismatch to some extent.

IV. CONCLUSION

The above report discusses various page ranking algorithms such as Page rank algorithm, HITS algorithm and semantic similarity algorithm. Each algorithm has different methods of ranking. A most common thing that all the algorithms use is the keywords based information retrieval. In future, the relevance can be improved by combining these techniques. Thus, page ranking enables search engines optimization.

References

[1] Nisha and Dr. Paramjeet Singh “A Review Paper on SEO based Ranking of web documents” International Journal of Research in Computer Science and Software Engineering Volume 4, Issue 7, July 2014.

[2] Mridula Batra and Sachin Sharma “Comparative Study of Page Rank Algorithm with Different Ranking Algorithms Adopted by Search Engine for Website Ranking” International Journal Computer Technology & Applications, Vol4 (1), 8-18.

[3] http://www14.informatik.tuenchen.de/konferenzen/Jass05/courses/6/Papers/07.pdf

[4] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit ” Comparative Study of HITS and Page Rank Link based Ranking Algorithms” International Journal of Advanced Research in Computer and Communication Engineering , Volume 3, Issue 2, February 2014.

[5] Parveen Rani and Er. Sukhpreet Singh, “An offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters” International Journal of Computers and Technology, Vol 9, No.1

[6] Yue He, Minghong Qiu, Maozhu Jin, Tao Xiong, “Improvement on HITS Algorithm” International Journal of Applied Mathematics and Information Journal, Vol 3, 1075-1086, 2012

[7] http://home.ie.cuhk.edu.hk/~wkshum/papers/pagerank.pdf

[8] http://pr.efactory.de

[9] http://www.google.com/technology/

[10] Deepti Kapila, Prof. Charanjit Singh, “Survey on Page Ranking Algorithms based on Digital Libraries” International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4,June 2014.