ISSN 22773061

An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters

Parveen Rani[1], Er. Sukhpreet Singh[2]
[1]Student, M.tech Final Year (CSE),
Guru Kashi University,
Talwandi Sabo (Bathinda), Punjab
parveenkamboj1989@gmail.com

[2]Assistant Professor, C.S.E Department,
Guru Kashi University,
Talwandi Sabo (Bathinda), Punjab
sukhpreet.manshahia@gmail.com

ABSTRACT

SEO stands for Search Engine Optimization. It is a technique that searches various web pages for specified keywords and ranks these Web pages according to some parameters. They are used to feed pages to search engines. The main importance of SEO is that it helps to find the relevant data and increase the rank of a webpage in search engine's results. In our paper, we develop a new algorithm M-HITS (Modified HITS) to provide the page rank. M-HITS Algorithm is a new version of HITS algorithm. It is developed by extending the properties of HITS algorithm.

Keywords: M-HITS, HITS, Hub, Authority.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 9, No 1
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

1. INTRODUCTION

Search Engine Optimization was a term used by web developers in the late 90s to highlight the importance of increasing a website's position in search engine's results. Search engine optimization (SEO) is a well defined and managed process which helps to increase the volume or improve quality of traffic to a web site from search engines. SEO activity helps to increase the amount of visitors to a Web site by ranking high in the search results of a search engine. The higher a Web site rank in the results of a search, greater the chance that site will be visited by a user. Search Engine Optimization involves the careful optimization of corporate web sites to effectively increase their visibility in the major search engines such as Google, Yahoo, Alta-Vista and many others. It makes the difference between a web site that has very little visibility and one that will be seen and found by millions of people. It is a program that searches documents for specified keywords and automatically fetches Web pages. Only meaningful results are returned for each query. They are used to feed pages to search engines. To locate any information from the web, the user accesses his favorite search engine, issues queries and clicks on the returned pages. The search results returned by a search engines are a mixture of large amount of relevant and irrelevant information. Any user cannot read all web pages returned in response to the user's query. Whenever you enter a query in a search engine and hit 'enter' button you get a list of web results that contain that query term. The web pages in such a setting are stored in different directories on the basis of their category. A search is generally based on the matches only in the descriptions submitted. Users normally tend to visit websites that are at the top of this list as they pick out those to be more relevant to the query. SEO is a technique which helps search engines find and rank your site higher than the millions of other sites in response to a search query. SEO thus helps you get traffic from search engines.

2. VARIOUS TECHNOLOGIES

2.1 Web Page Crawling:

A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier**.** The high rate of change implies that the pages might have already been updated or even deleted**.** First, search engines crawl the Web to see what is there. This task is performed by a piece of software, called a crawler or a spider**.**

2.2 Spider:

Spiders are used to feed pages to search engines. It's called a spider because it crawls over the Web. Most Web pages contain links to other pages, a spider can start almost anywhere. As soon as it sees a link to another page, it goes off and fetches it. Large search engines, like Alta Vista, have many spiders working in parallel. A spider visits Web sites and reads their pages and other information in order to create entries for a search engine index.

3. PROPERTIES OF WEB PAGES

3.1 Hub: Pages that contain links to authorities. Hubs are pages that act as resource lists, guiding users to authorities. A better hub points to many good authorities.

3.2 Authority: Pages that provide important information. A valuable and informative webpage usually pointed to by a large number of hyperlinks. Authorities are pages having important contents.

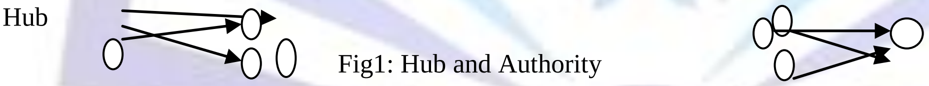


Fig1: Hub and Authority

3.3 Bold, Italic and Specific Words:

Some special words which are written in bold and italic form. For example: Headings, some specific words (e.g. Computer, CPU) etc.

3.4 Number of Unique Clicks:

When we search any word in the Google search engine, there will be opened many links. Some links contain most important information so that we click many times on the same link. How many times we will click on that link; the Google search engine will count the number of clicks individually.

4. HITS ALGORITHM

HITS algorithm is proposed by Kleinberg in 1999. Kleinberg proposed the hypertext induced topic search algorithm for topic search on the WWW. HITS algorithm is also developed for ranking documents based on the link information among a set of documents. Brin and Page proposed two level schemes for importance of web pages hub identity and authority identity. The HITS algorithm starts with a focused hyperlink graph for a WWW for a query. It then does iterative, eigenvector based computation to identify good hub pages and authority pages. The hub identity captures the quality of the page as a pointer to useful resources, and the authority identity captures the quality of the page as a resource itself. A good authority is a source of useful information, while a good hub is a page that contains a useful collection of links.

5. PURPOSED METHOD: M-HITS

Basic: Our approach ranks the web pages according to the following parameters of a web page:

There are used six parameters:

1. Bold words
2. Italic words
3. Keywords
4. Number of unique clicks
5. Hub Values
6. Authority Values

In our approach we have assigned various weights to the above keywords. For example we assign a value of ".05" to a bold keyword. Which means in a specific keyword is found in the webpage in bold then it increases the rank by ".05" and so on. Here is a list of parameters and there corresponding values for calculation of webpage Rank:

Sr. No.	Parameter Name	Value
1	Bold Keyword	.05
2	Italic Keywords	.02
3	Keywords	.02
4	No. of unique clicks	.01
5	Hub Values	As calculated
6	Authority Values	As calculated

Table1. List of parameters

6. M-HITS Algorithm:

Following steps are used:

Step1. Input adjacency matrix of web pages

Step2. Input the frequency of various types of parameter (Bold, Italic, Keyword, and No. of Unique Click).

Step3. Calculate the Hubs and Authorities for each web page.

Step4. Normalize these values for each web page and calculate the partial rank for each web page.

Step5. Add weights of parameters to the calculated partial ranks.

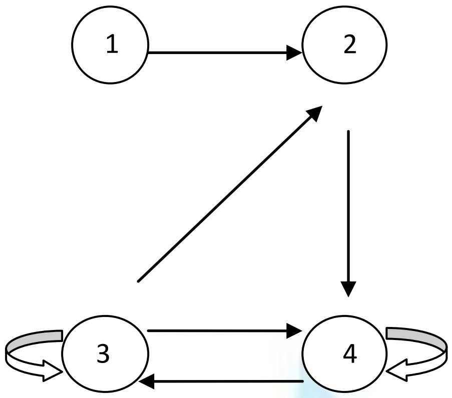
Step6. Sort the web pages positions according to the calculated ranks corresponding to both Hub Values and Authority Values of web pages.

Step7. Exit

7. IMPLEMENTATION AND RESULT

7.1 Implementation of M-HITS algorithm:

7.1.1 Take a Directed graph which is showing the link between 4 pages. There are 4 pages; 1,2,3,4.



7.1.2 Make a 4*4 matrix from this graph

$$\begin{array}{ll} & \text{Authority} \\ & \downarrow \\ \text{Hub} \rightarrow & 0\ 1\ 0\ 0 \\ & 0\ 0\ 0\ 1 \\ & 0\ 1\ 1\ 1 \\ & 0\ 0\ 1\ 1 \end{array}$$

7.1.3 Count the number of Hubs and Authorities:

Min hub = 1	Min Authority = 0
Max hub =3	Max Authority = 3

7.1.4 Then normalized all hubs and Authority:

Norm hub= 0,	Norm Authority = 0
Norm hub= 0,	Norm Authority = 0.6666667
Norm hub= 1,	Norm Authority = 0.6666667
Norm hub = 0.5,	Norm Authority = 1

7.1.5 Parameters used: weight age for B= 0.05, I= 0.02, S= 0.02, C= 0.01

Bold (B)	**Italic (I)**	**Specific word (S)**	**No. of unique clicks (C)**
1	2	1	4
2	1	2	5
32	5	2	1
1	6	3	9

Table2. Parameters Used

7.1.6 Normalized hubs and authorities according to the parameters:

Norm hub= 0.15, Norm Authority = 0.15

Norm hub= 0.21, Norm Authority= 0.87667

Norm hub= 2.75, Norm Authority = 2.4167

Norm hub = 0.82, Norm Authority = 1.32

7.2 Results of M-HITS algorithm:

7.2.1 Page rank According to Hub:

1) Page rank = 1 Page Number = 3 Page Value = 2.75

2) Page rank = 2 Page Number = 4 Page Value = 0.82

3) Page rank = 3 Page Number = 2 Page Value = 0.21

4) Page rank = 4 Page Number = 1 Page Value = 0.15

7.2.2 Page rank According to Authority:

1) Page rank = 1 Page Number = 3 Page Value = 2.4167

2) Page rank = 2 Page Number = 4 Page Value = 1.32

3) Page rank = 3 Page Number = 2 Page Value = 0.8766

4) Page rank = 4 Page Number = 1 Page Value = 0.15

8. CONCLUSION AND FUTURE WORK:

Webpage rank plays an important part to search the documents for the specific keywords. When a person query a search engine to find the web pages according to his requirement then search engine generated thousands of web pages. But the main problem of user is that it can’t access all these documents and also all these documents doest not contain relevant information. So it is a responsibility of a search engine to sort these web pages according to web page rank so that user can find the results quickly. Web page ranks are assigned to web pages by a SEO program. More the parameters a SEO program uses to rank the web page more accurately a rank can be evaluated. In the previous work, HITS Algorithm was implemented for Webpage Rank. In HITS algorithm; firstly; by using directed graph, an adjacency matrix had made and then calculate all the Hubs values and Authorities values. From all the Hubs values and Authorities values, HITS algorithm was found the Min Hub, max Hub, Min Authority and Max Authority. The Value of Min Hub, Max Hub, Min Authority and Max Authority were the result in HITS Algorithm. There are two parameters used in HITS Algorithm. In our system we have used 6 parameters to evaluate the rank for a web page.

In future, we can also use some AI techniques in addition to these proposed techniques to improve the rank of web pages.

REFERENCES

[1] Nidhi Grover and Ritika Wason,“Comparative Analysis of Page rank And HITS Algorithms”, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October 2012,ISSN: 2278-0181.

[2] Gyanendra Kumar, Neelam Duhan, A. K. Sharma, “Page Ranking Based on Number of Visits of Links of Web Page”, International Conference on Computer & Communication Technology (ICCCT)-2011 IEEE.

[3] Harmunish Taneja , Richa Gupta, “Web Information Retrieval using query independent Page Rank Algorithm”, 2010 International Conference on Advances in Computer Engineering, 2010 IEEE.

[4] Mr.Ramesh Prajapati, “A Survey Paper on Hyperlink-Induced Topic Search (HITS) Algorithms for Web Mining”, International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 2, April – 2012, ISSN: 2278-0181.

[5] Page Ranking Algorithms for Web Mining, “Rekha Jain, Dr. G. N. Purohit”, International Journal of Computer Applications (0975 – 8887), Volume 13– No.5, January 2011.

Ms. Parveen Rani, I have received my B-Tech degree in computer Engineering from Yadavindra College of Engineering, Talwandi Sabo (Bathinda) in 2011 and pursuing M-Tech in computer science & Engineering from Guru Kashi University, Talwandi Sabo (Bathinda). My Research areas are Internet technology and Web Mining.

Er. Sukhpreet Singh, Assistant Professor, CSE Department, Guru Kashi University, Talwandi Sabo.

Qualification: B-Tech (C.S.E) from Chitkara Institute of Engineering & Technology, Rajpura, Patiala (Punjab). M-Tech (Comp. Engg.) from Punjabi University Patiala (Punjab).