Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review Paper on SEO based Ranking of Web Documents

Nisha
Dept. of Computer Science & Engineering
GZS PTU Campus,
Bathinda, India

Dr. Paramjeet singh
Dept. of Computer Science & Engineering
GZS PTU Campus,
Bathinda, India

Abstract— With the use of web search engines everyone's life becomes more convenient. But also with the daily growth in web, it becoming very difficult to get the proper information related to user search query. When user searches some information, search engines return huge amount of web pages in response to search queries. It is not possible to one and all to explore all web pages. So there is need of such search engine that could provide useful and quality information to internet user. To do so, there is a technique called SEO i.e. search engine optimization. SEO is a process that uses search principles for search engines to provide the higher ranking to more suitable webpage. It is done by applying various webpage ranking algorithm. The aim of this paper is to cover techniques for optimizing the search engines which are helpful for internet user to make available valuable data.

Keywords— Web-page Ranking, SEO Techniques, Semantic Similarity, HITS, Domain Dictionary

I. INTRODUCTION

Search Engine Optimization (SEO) is a technique that refers to process of improving traffic to website by increasing the visibility of site in search engine results. SEO helps search engine to give highest ranking to a particular website so that it could be shown at the top of search engine result page in response to user search query. And the searched data must be relevant to user query. If a website get place at top of results then there are more chances of user visiting that web site. To do so websites uses no. of optimization methods. Similarly semantic similarity is also important. Semantic similarity is one that fulfills the user expectations. As it is not only sufficient to search information on the bases of keyword matching that are presented on the web. On web, there are so many documents that have same information regarding user query but have different keywords. To overcome this problem, semantic similarity is must.

There are some terms that used in web ranking.

1) *Web Search Engine:* Web search engines are program that return lots of web pages in response to user query. It is kind of term that are used to describe systems like Google, Bing and Yahoo! Etc.
2) *Spider:* Spider is a kind of program that is used by search engines to supply pages to search engines. Spider studies the content on collection of web pages and records any hyperlinks it finds. Then search engine spider follows these URLs, collects all the data by saving copies of the web pages.
3) *Indexer:* Indexer is another kind of program that read the data collected by spider and creates an "Index" of collected data for use of visitors. Search engine use index for quickly providing search results**.**
4) *Crawler:* Crawler is programs that visit the websites which are provided to it. While spider visit to its further links also.

Advantages:

1) *Useful for web-sites owners:* Webpage ranking helps for increasing web-site traffic that will leads to more customer. As it Increase brand name or product visibility and also raise the business services and also Increase in sales and popularity. In actual fact, Organic listings are free. When web pages are listed at the top, then there is no need to pay per click or give out a budget for advertising.
2) *Helpful for Internet users:* When internet user trying to get information then if efficient SEO technique and ranking algorithms was used by search engine then user can get relevant data corresponding to user query. User can get information in between few pages of search so there is no need to explore whole search result.

II. LITERATURE SURVEY

[1] P. Chahal, M. Singh and S. Kumar "Ranking of Web Documents using Semantic Similarity"

This paper proposed a novel technique which makes user search data quite efficient. This technique gives a relationship or similarity between searched document and user query. It is also consider the semantic structure of document and user query.

The result set obtained from this approach gives better results than prevailing approaches. The future work can be done by using deeply semantic analysis of web pages and relevance of documents.

[2] G. Kumar, N. Duhan and A. K. Sharma "Page Ranking Based on Number of Visits of Links of Web Page"

In this paper author presented a modified page ranking algorithms which is more target oriented than original page rank. The modified algorithm calculates page rank value or importance of web pages based on the visits of incoming links on a page. The paper presented a novel page ranking algorithm called VOL that provides more relevant results than original Page Rank. As a result, Author proved that VOL is far dynamic than original Page Rank algorithm and also observed that the page which has more visits of incoming links is carrying more rank value than less visited pages. The paper also presents a method to find link-visit counts of Web pages and a comparison between VOL with the Page Rank algorithm.

[3] P. Rani and Er. S. Singh, "An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters"

This paper describes the new algorithm for calculating web page rank according to different parameters. The proposed algorithm called M-HITS (Modified HITS) is a new version of HITS algorithm. It is developed by extending the assets of HITS algorithm. Author present new algorithm in which six parameters are used to evaluate rank for web page. Future work can be done by using some AI techniques in addition to these proposed techniques to improve the rank of web pages.

[4] A. Jain, R. Sharma, G. Dixit and V. Tomar "Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages"

This paper proposed a new method called Intelligent Search Method (ISM). Author developed new method to index the web pages using an intelligent search strategy in which meaning of the search query is interpreted and then indexed the web pages based on the interpretation. This Paper also described the limitations of existing methods and discussed the different algorithms used for link analysis like Page Rank (PR), Weighted Page Rank (WPR), Hyperlink-Induced Topic Search (HITS) and CLEVER algorithm. The new method can be integrated with any of the Page Ranking Algorithms to produce better and relevant search results.

[5] N. Batra, A. Kumar, Dr. D. Singh and Dr. R.N. Rajotia "Content Based Hidden web Ranking Algorithm (CHWRA)"

This paper proposed the algorithm called CONTENT BASED HIDDEN WEB RANKING ALGORITHM (CHWRA). The Proposed ranking algorithm consists of four different attributes. This method tries to cover all the aspects which directly or indirectly affect the popularity of the web page. This method is an effort to generate an ordered Hidden web searched result set. The ranking algorithm (CHWRA) gave the desired result. Further the work that can be done will be based on parameter i.e. *Access Time length* which can be used as "Duration of time that spent on a web page" that means the total time that user access particular web page.

III. EXISTING TECHNIQUES

Following are some most commonly used search engine optimization techniques.

3.1 Page Rank Algorithm: Page Ranking Algorithm is Surgey Brin and Larry Page. Page ranking algorithm is based on link structure of web. Page ranking algorithm is used by the popular search engines to rank the web page. It basically works on link structure on web pages. This algorithm counts the no. of links in web page and estimate the importance of web site. According to Page Rank Algorithm, The web site will be more important if it have more no. of links. The algorithm assigns a numerical weight to each element of World Wide Web so it referred to as page rank of E and it is denoted by PR (E).x

3.2 HITS Algorithm*:* HITS Algorithm stands for Hyperlink-Induced Topic Search Algorithm. This algorithm is developed by Jon Kleinberg. It is also called Link Analysis algorithm. HITS algorithm is used to rank the web page focusing on *Hubs and Authority*. When user issue some search query HITS algorithm expands list of web pages returned by search engines in response to user search query. Hits algorithm is query dependent. The famous Twitter web site uses the HITS style algorithm.
HUB: Hub represent the page that point to the authorities.
Authority: Authority represent as a source of valuable information.

Steps of HITS Algorithm:

Step 1: Enter the adjacency matrix of the web pages.
Step 2: Then enter the frequency of different types of parameters (Bold, Italic, Keyword, and No. of Unique Click).
Step 3: Compute the Hubs and Authorities for each web page.
Step 4: Normalize all these values for every web page and then compute the partial rank for every web page.
Step 5: And then add weights of the parameters to the calculated partial rank.

Step 6: Sort the web pages positions according to the calculated ranks corresponding to both Hub Values and Authority Values of web pages.
Step 7: Exit.

3.3 Semantic Similarity: Semantic similarity concept used in many fields. When the data or documents have same meaning then semantic similarity approach is used based on identical keywords which are extracted from different documents.

Steps of semantic similarity Algorithm:
Step1: Firstly construct a Text-List (by links).
Step 2: Then acquire query as a text: a String.
Step 3: For each Text in Text-List do:
(a) Create Text-Vector-Space.
(b) Create Domain-Dictionary of words.
(c) Using Statistical-Model () and Domain-Dictionary,
Compute relevance-value of Text corresponding to user Query.
(d) Make Domain-Ontology of the Text.
(e) Compute Domain-Similarity of Text value with
Domain- Ontology.
(f) Verify the maximum of Domain-Similarity value and relevance-value and call it Relevance-Score.
Step 4: Then Go to step 3 until there are no text left in the Text-List or else no more text is to be considered.
Step 5: Organize the links according to the decreasing order of relevance-score and assign the rank to them.
Step 6: And then finally display the contents according to their ranks.

IV. RESULTS & DISSCUSION

Implementation: Semantic Similarity algorithm has been implemented by using C# programming language and Visual Studio 2010.
Semantic Similarity:
Snapshot 1:

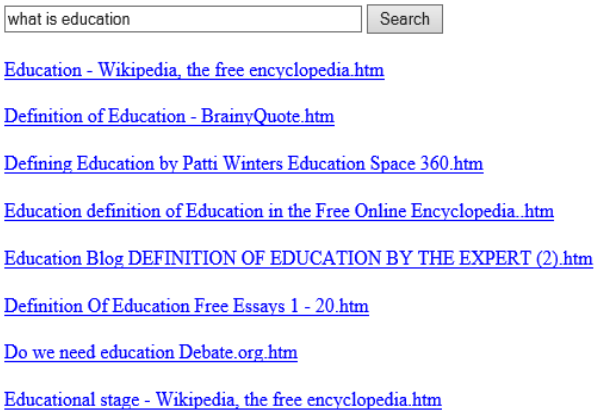

Figure 1: Output of Semantic Similarity

The above figure represents the practical implementation of existing semantic similarity. The result is produced according to rank which is calculated through dictionary which is created manually. In this screen shot, when query "what is education" has been entered then it produced set of web pages in response to the user query. In Table I, actual rank is score of web pages given by human and variance values are calculated.

TABLE I: RANK OF WEB PAGES

S. No.	Actual Rank	Google Rank	Variance by Google Rank	Our Rank	Variance by Our Rank
1.	D1,D2,D4,D3	D2,D4,D3,D1	10	D1,D2,D3 ,D4	2
2.	D21,D23,D25,D 26,D22,D24	D21,D22,D23, D24,D25,D26	34	D21,D25, D26,D22, D23,24	10
3.	D32,D33,D31,D 34	D32,D33,D34, D31	18	D31,D32 D33,D34	6

Hits Algorithm:

The following figure shows the result of HITS algorithm. In this algorithm hubs and authority values are calculated and finally normalized all values of hubs and authority. In figure 2 all the details of web pages are manually entered and then Min. Hub and Max Hub, Min Authority and Max Authority is calculated.

Snapshot 2:

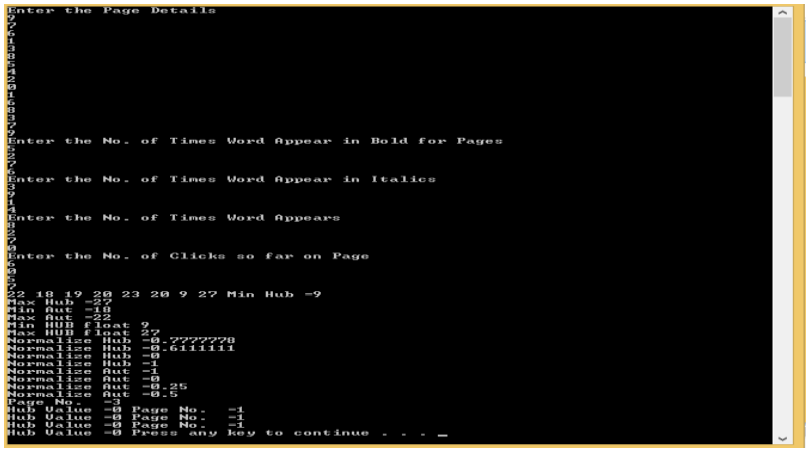

Figure 2: Output of HITS Algorithm

Parameter Definition:

- *Time efficiency:* It is define as the avg. time require for searching the web pages from a set of database. Time efficiency is calculated as follow:

  Time Efficiency = Avg. time to search web page/Total No. of web pages *100
- *Accuracy:* Accuracy is whether the page with highest page rank found on top or not i.e. system display web pages according to the web page rank calculated.
- *User Specific Page Generation:* This parameter specify that whether the page display is according to user interest or not.
- *Relevance Ratio:* It specifies that how much percent of content is relevant to the input query.
- *High Relevance Ratio:* It is the difference between the higher ranked web page and the lower ranked web page.

Comparison of the Existing Techniques:

TABLE II: COMPARISON OF TECHNIQUES

Parameter/Technique	HITS Algorithm	Semantic Similarity Algorithm
Time Efficiency	72%	87%
Accuracy	79%	91%
User specific Page Generation	No	No
Relevance Ratio	90%	92%
High Relevance Ratio	30%	41%

Graph:

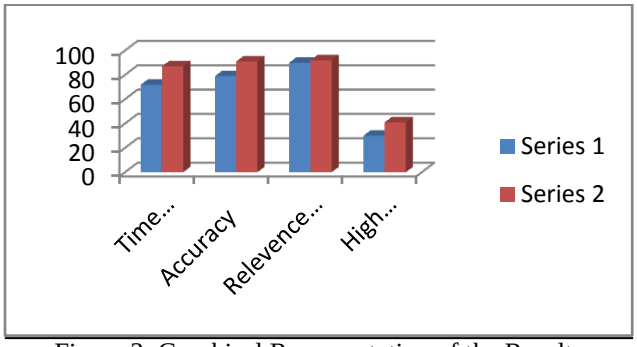

Figure 3: Graphical Representation of the Results

V. CONCLUSION

In this paper, authors present the review on search engine optimization techniques. Web ranking algorithms such as Hits algorithm and Semantic similarity algorithm are implemented in order to compare these techniques and analyze their results mainly in terms of web ranking. Authors concluded that there is need to improvise existing techniques so that better results can be achieved. The proposed method will be result in having more efficiency and more relevance web pages in response to user query.

REFERENCES

[1] P. Chahal, M. Singh and S. Kumar "*Ranking of Web Documents using Semantic Similarity*" Information Systems and Computer Networks (ISCON), 2013 International Conference on 2013 IEEE, DOI 10.1109/ICISCON.2013.6524191 Page(s): 145 - 150

[2] G. Kumar, N. Duhan and A. K. Sharma "*Page Ranking Based on Number of Visits of Links of Web Page*" Computer and Communication Technology (ICCCT), 2011 2nd International Conference on 2011 IEEE, DOI 10.1109/ICCCT.2011.6075206 Page(s): 11 – 14

[3] P. Rani and Er. S. Singh, "*An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters*" INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY Vol. 9, No 1 J u l y 1 5, 2 0 1 3

[4] A. Jain, R. Sharma, G. Dixit and V. Tomar, "*Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages*" Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013 IEEE, DOI 10.1109/CSNT.2013.137 Page(s): 640 – 645

[5] N. Batra, A. Kumar, Dr. D. Singh and Dr. R.N. Rajotia "*Content Based Hidden web Ranking Algorithm(CHWRA)*" Advance Computing Conference (IACC), 2014 IEEE International, 2014 IEEE, DOI 10.1109/IAdCC.2014.6779390 Page(s): 586 – 589

[6] H. Dubey, Prof. B. N. Roy "*An Improved Page Rank Algorithm based on Optimized Normalization Technique*" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5), 2011, Pages(s): 2183-2188

[7] R. Kumar and S. Saini "*A Study on SEO Monitoring System Based on Corporate Website Development*" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.2, June 2011

[8] K. ur Rehman and M. N. Ahmed Khan " *The Foremost Guidelines for Achieving Higher Ranking in Search Results through Search Engine Optimization*" International Journal of Advanced Science and Technology Vol. 52, March, 2013

[9] N. V. Pardakhe, Prof. R. R. Keole "*Analysis of Various Web Page Ranking Algorithms in Web Structure Mining*" International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013

[10] M. Cui, S. Hu "*Search Engine Optimization Research for Website Promotion*" Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on Vol.4, 2011 IEEE, DOI 10.1109/ICM.2011.308 Page(s): 100 - 103

