A Ranking Of Web Documents Using Semantic Similarity And Artificial Intelligence Based Search Engine

*Seema Rani*1 , *Upasana Garg2*

1Guru Kashi University, Department of CSE,
Talwandi Sabo, Punjab, India

2Guru Kashi University, Department of CSE,
Talwandi Sabo, Punjab, India

Abstract: **People generally access the information over the internet with the help of search engines. Search engines are the programs which find the specific pages for users according to their query. Web page ranking is the most important factor on internet for search engines. Web page ranking is a technique that ranks the web pages according to their different qualities and parameters for search engines. There are various web search engines are available on internet some of them are Google, Yahoo, and Bing etc. In this paper, we present a new web ranking system by using Semantic Similarity and HITS algorithm along with AI technique. These techniques work together to rank a web page from a number of web pages on the internet.**

Keywords: **SEO, Search engine Optimization, Web page ranking algorithm,HITS algorithm, Semantic Similarity Algorithm**

I. INTRODUCTION

Now a days searching on the internet is most widely used operation on the World Wide Web. The amount of information is increasing day by day rapidly that creates the challenge for information retrieval. There are so many tools for perform efficient searching. Due to the size of web and requirements of users creates the challenge for search engine page ranking. Web page Ranking is the main part of any information retrieval system. In the Search Engine

1. Crawler: used for retrieves the web pages and web contents

2. Indexer: stores and indexes information on the retrieved pages

3. Ranker: Measure the importance of Web Page

4. Retrieval Engine: performs lookups on index tables against query

The web has a hierarchical structure: every day pages are added, deleted, and modified. The size of the web is on the order of more than a billion pages, and many of those pages contain redundant or incorrect information. A search tool on the web must be able to distinguish high-quality pages from low-quality pages. In addition, users of web search tools also present a challenge to IR researchers and developers. The average web IR user enters very short queries, does not make use of system feedback to revise the query, seldom performs a search using advanced search options, and generally views only the top few documents returned by the search If user sends the query for particular topic, then Web can have hundred, even thousands results regarding that query. But if the Ranking algorithm does not provide the result within the top few positions of the ranking then that search engine is useless and not efficient. The users have no patience to go through the hundred pages to find the one which they want. So the quality ranking of web page becomes essential. The needs of users are different, so random page may be highly relevant to the query . The main task of ranking of web page is to identify the importance of web page. Mainly In links to the pages and out links from the page can give idea about the context of the page. In this thesis we will discuss three algorithms for Ranking of Web Pages which is Semantic Similarity, HITS and Artificial Intelligence. Internet provides huge amount of information that people access daily with the help of search engines sometimes search engines produced results are not according to requirement of user. There are many results produced by search engines and some of them are not associated to user query. This is a reason that there is a need for some kind of techniques or system that can help user to get rid of this problem. Search engines

explore websites contents to gather information about a website. Therefore, there is a need to optimize a website to make it search engine friendly. Web search engines are considered as intermediate between user and information repository. Search engines used Crawler, spider, and indexer programs to present web pages. Crawler visits the web sites that are provided to it while spider visits to its further links also.

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] Gyanendra Kumar, Neelam 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] Parveen Rani and Er. Sukhpreet 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 properties 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] Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha 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.

III. PROPOSED WORK

The proposed system rank the webpage according to three techniques which are Semantic Similarity approach, HITS as well as on the basis of AI technique which access the user history to rank the webpage according to the user query. An online interface is developed using asp.net web technologies and c#.net is being used as an programming language tool. A Graphical User Interface (GUI) will be created to display the results.

Algorithm of Proposed system has the following steps:

Steps of semantic similarity Algorithm along with AI technique:

Step1: Firstly construct a Text-List (by links).

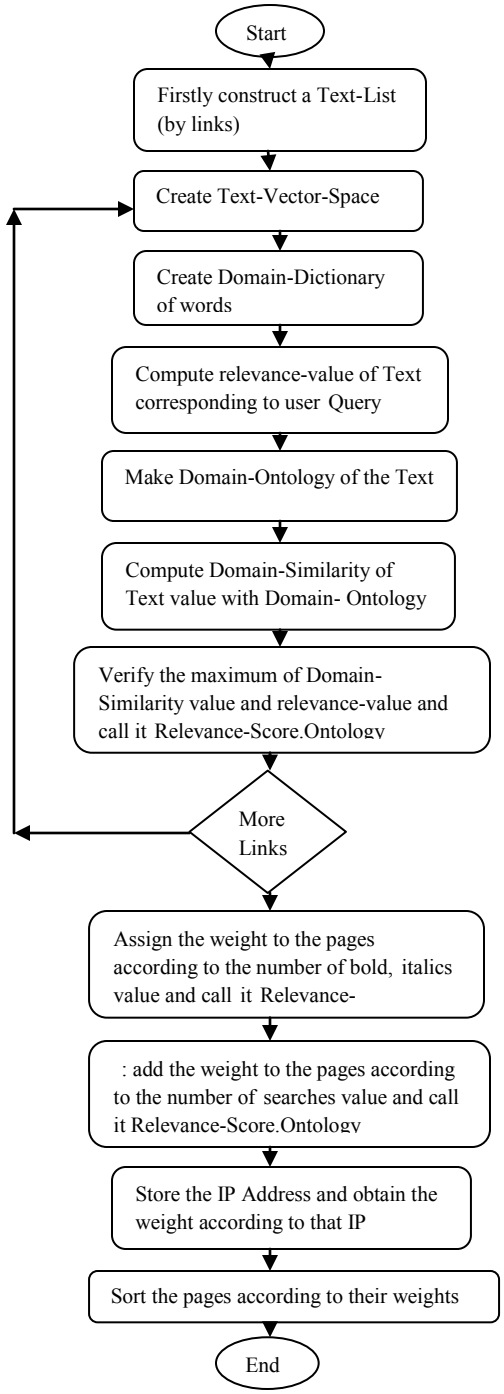
Step 2: Then acquire query as a text: a String.

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: Assign the weight to the pages according to the number of bold, italics keywords present according to the user query.
Step 6: add the weight to the pages according to the number of searches performed by the user for that type of query.
Step 7: Organize the links according to the decreasing order of relevance-score and assign the rank to them.
Step 8: And then finally display the contents according to their ranks.



IV. RESULTS AND DISCUSSION

The proposed system is tested on the various input queries collected from the various people. The overall accuracy of the proposed system is evaluated to 95% which is shows very good results. In the proposed system, web page ranking is performed according to the user specific search. Proposed system contain 1000 page entries from which system find the most appropriate page according to the user requirement.

Parameter\Technique	HITS Algorithm	Semantic Similarity Algorithm	Proposed System
Time Efficiency	72%	87%	91%
Accuracy	79%	91%	95%
User specific Page Generation	No	No	Yes
Relevance Ratio	90%	92%	96%
High Relevance Ratio	30%	41%	51%

Result Table of proposed system :

V. CONCLUSION AND FUTURE SCOPE

Proposed system presents a improved Semantic Similarity technique to rank a web page from a set o given web pages. System is using semantic similarity algorithm along with AI technique to rank the WebPages. System is tested on 1000 web pages comes under various categories like education, computer, programming, chemistry etc. Various input queries are given as an input to the system and results are checked. Overall accuracy of the system evaluated to 95%. System can be further improved by implementing on the cloud servers and by using multithreading techniques to improve the time efficiency. System can be further checked by increasing the number of web page categories. Multithreading techniques can also be integrated in the proposed system to improve the overall performance of the system.

VI. REFRENCES

[1] P. Chahal, M. Singh and S. Kumar "Ranking of Web Documents using Semantic Similarity" *2013 International Conference on 2013 IEEE, Page(s): 145 - 150*

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

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

[4] Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar "Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages"

[5] Dr. Khanna SamratVivekanand Omprakash "Concept of Search Engine Optimization in Web Search Engine"

[6] Laxmi Choudhary and Bhawani Shankar Burdak "Role of Ranking Algorithms for Information Retrieval"

[7] Joeran Beel , Bela Gipp and Erik Wilde "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co."

[8] Chris H.Q. Dingy, Hongyuan Zhaz , Xiaofeng Hey , Parry Husbandsy and Horst D. Simony " Link analysis: Hubs and Authorties on the world wide web"

[9] Jaideep Shrivastva, Prasanna Desikan and Vipin Kumar".

"Web Mining - Concepts, Applications & Research Directions"

[10] Mr.Ramesh Prajapati "A Survey Paper on Hyperlink-Induced Topic Search (HITS) Algorithms for Web Mining"

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

[12] 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)

[13] 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, 2183-2188*

[14] 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*

[15]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*

[16] Bussa V.R.R. Nagarjuna, Akula Ratna babu, Miriyala Markandeyulu, A.S.K.Ratnam "A web mining : algorithms methodology and application"

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

[18] M. Cui, S. Hu "Serach Engine Optmization Research for Website Promotion" *information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on (Volume:4) 2011 IEEE, Page(s): 100 – 103*

[19] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, & Panayiotis Tsaparas, ―Finding Authorities and Hubs from link structures on the World Wide Web‖ , *in Proceedings of the 10th WWW Conference, Hong Kong, 2001, pp. 415-429.*

[20] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, & Panayiotis Tsaparas, ―Link analysis ranking: algorithms, theory, and experiments‖ , *in ACM Trans. Inter. Tech., 5(1) , 2005, pp. 231-297.*

Author Profile

Seema Rani received the B.Tech degree in Computer Engineering from Punjab technical University Jalandhar in 2012. Currently She is pursuing M.Tech degree in Computer Science from Guru Kashi University, Talwandi Sabo, Bathinda (Punjab). Her research interests Internet Technology
And Web Mining.

Er.Upasana Garg, Assistant Professor, CSE Department, Guru Kashi University, Talwandi Sabo. Qualification: B-Tech (C.S.E) from JCDM College of Engineering & Technology (Kurkshetra University) M-Tech (Comp. Engg.) from CDLU Sirsa. Her Reasearch area is Computer Networking.

