Improved Page Ranking in Search Engine Using Web Content Mining and Web Structure Mining

Sapna Kaushik¹, Babita Yadav²

¹M.tech Scholar, MVN University Palwal, Haryana, India
²Deptment of Computer Science, MVN University Palwal, Haryana, India

Abstract—Search Engine is a common source of information in this era of Internet. People explore the web using search engine for their information need. Search engine depends upon page rank algorithm to return a relevance result set of web pages for a user query. The performance of page ranking system is an area of research for many researchers in recent years. Web structure mining and web content mining are two major categories of page rank algorithms available in literature. Both of these techniques are having their advantages and disadvantages. In this paper a new approach for page ranking is proposed. It uses a combined technique of web structure mining and web content mining for page ranking. The algorithm is implemented and its performance is compared with other existing algorithms. The results of implementation validate that the proposed algorithm works better than the existing algorithms.

Keywords— Web structure mining, web content mining page rank, weighted page rank.

I. INTRODUCTION

In today's world search engines such as Google, Bing are major source of information. Search engines rely on performance of page rank algorithms to provide relevant results for user queries. Google’s Page Rank algorithm is one of the best-known algorithms in web search. With the increasing number of Web pages and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines need to be more efficient in its processing way and its output. The more popular web pages are, the more linkages that other web pages tend to have to them or are linked by them. The proposed extended Page Rank algorithm—Weighted Page Rank Algorithm assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The page ranking is an algorithm developed by Sergey Brin and Lawrence Page in 1998 and used by Google search engine. The algorithm assigns a numerical value to each element in the world wide web for the purpose of measuring the relative importance of each page, the idea being to give a higher page rank value to a page that is frequently visited by users..

II. LITERATURE SURVEY

Madhurdeep Kaur and Prof Charanjit Singh [1] presented a novel hybrid page rank algorithm based on both content and link structure of web pages. A set of repository of web pages in the database are selected. A link structure will be created that will explain how these pages are linked to each other. It uses a concept of similarity of content and query. Accordingly, both the page rank and content rank will be compared. Pijitra Jomsri and Dulyawit Prangchumpol [2] devised and proposed a hybrid model ranking for research paper searching and bookmarking. The keywords used for ranking are the year of the published paper and type of research paper this is combined with the similarity ranking. The proposed approach uses the concept of indexing using tf-idf. Sharmila Subudhi [3] proposed the hybrid page ranking algorithm which is based on prediction and calculation of different numbers of backlinks to a particular web page. It uses a concept of coordinate list and compressed sparse row which calculates the number of backlinks to a page using sparse matrices.
Lissa Rodrigues and Shree jaiswal [4] proposed a hybrid approach of enhanced ratio rank and page level keyword algorithm. This approach uses the combination of link based mining and page level keyword search. In enhanced ratio rank both inlinks and outlinks are considered along with the times the user visits a particular link. Veningston, Dr R Shanmugalakshmi[5] proposed a enhanced web search re-ranking algorithm by incorporating user profile information obtained from various sources .this paper also discusses the problems of present web search.

III. PROPOSED WORK

In this paper a new improved approach using web structure mining and web content mining is proposed. Content-based ranking algorithm In this approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and link structure provided by search engines. Based on the user query, search engine results are retrieved. Every result is individually analyzed and interpreted based on keywords and content. Web Structure based ranking algorithms The link analysis algorithm is based on link structure of documents. The quality of the results from search engines are generally lower in quality than what user expects and quality can be improved greatly if the pages are ranked according to a useful criteria based on the links between the pages ie each page may have the reference to other set of pages. They view the web as a directed graph where the web pages form the nodes and the hyperlinks between the web pages form the directed edges between the nodes. There are two famous link analysis methods

In this paper a new page ranking using web structure mining and web content mining is proposed. The algorithm work as follows:

Algorithm : Algorithm for Page Ranking using Web Structure Mining and Web Content Mining
Input : A list of web pages
Output : A ranked list of web pages for a user query
1. Read all the web pages and calculate the page rank, weighted page rank of these web pages using Page Rank and Weighted Page Rank algorithms.
2. Read a user query and find the web content page rank of the web pages using following steps
   2.1 Find the frequency of query word(s) in all the web pages
   2.2 Sort the web pages according to decreasing order of the frequency of the query word(s)
   2.3 The sorted sequence will give the content page rank of the documents for the user query
3. Find the average rank of the documents using web structure rank found in step-1 and web content rank found in step-2
4. Sort the documents in descending order on the average rank found in step-3
5. Return sorted sequence of the documents in step-4 as the final ranked list

IV. RESULT AND ANALYSIS

The proposed algorithm is implemented in JAVA using JDK7 and NetBeans IDE. A sample of 10 web pages is taken as a data set. The results can be compared using well-known metrics for analyzing the performance of page ranking algorithms which is precision.

Precision
Precision is a measure to find the fraction or retrieved documents which are relevant to user query. It is used to calculate the quality of results retrieved for a user query from a search engine. It is calculated as follows :

\[
\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}
\]
Precision = (|\{Relevant Documents\} \cap \{Retrieved Documents\}|) / |\{Retrieved Documents\}|

We calculate the value of precision for existing algorithm. We enter a query ‘UNIVAC’ and we get 10 web pages for that. Out of these 10 web pages only 8 web pages are relevant.

Query Word : UNIVAC
Number of documents retrieved = 10
Number of relevant documents retrieved = 8
Precision = 8/10 = 0.8

Calculation of value of precision for proposed algorithm
We enter same query ‘UNIVAC’ and we get 8 web pages for that. All these 8 web pages are relevant to user query.
Number of documents retrieved = 8
Number of relevant documents = 8
Precision = 8/8 = 1.0

The proposed algorithm improves the value of precision. For existing algorithm the value of precision is 0.8 and for proposed algorithm value of precision is 1.0. So the proposed algorithm improves the value of precision by 20%. Figure-1 show a bar chart to compare the value of precision for existing algorithm and proposed algorithm.

![Figure 1](image-url)

**Figure 1** Comparison of value of precision for existing algorithm and proposed algorithm

V. CONCLUSION AND FUTURE SCOPE

This paper propose and implemented a page rank algorithm. The algorithm take advantage of web structure mining and web content mining algorithms. The results of the proposed algorithm is better than the existing weighted page rank algorithm. The proposed algorithm improves the value of precision by 20% as compared to existing algorithm. But this work has some limitations and here are some directions for future work:
1. The work can be implemented and tested on a large number of web pages
2. The proposed algorithm can be implemented on a live data set
3. The results can be compared with all other existing page rank algorithms such as hit-based algorithms, context aware page rank algorithm etc.

REFERENCES

[2] Pijistra Jomsri and Dulyawat Prangchumpol A hy brid model ranking search result for research paper searching on social bookmarking in 2013 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)
[3] Sharmila Subudhi in Designing a hybrid page ranking algorithm for semantic web search engine in 2013 International journal of scientific research(jsr)
[5] Veningston K and Dr R.Shanmugalakshmi Enhancing personalized web search re-ranking algorithm by incorporating user profile published in IEEE 2013
[6] Hema Dubey and BN Roy An Improved Page rank Algorithm Based on optimized Normalization technique in IJCSIT 2013