An Effective Content Based Web Page Ranking Approach

NIDHI SHALYA
M.Tech (Computer Science),
Amity School of Engineering & Technology,
Amity University, Lucknow, India 226010
nidz.cs@hotmail.com

SHASHWAT SHUKLA
Computer Science Department,
Amity School of Engineering & Technology,
Amity University, Lucknow, India 226010
shashwatshukla10aug@gmail.com

DEEPAK ARORA
Computer Science Department,
Amity School of Engineering & Technology,
Amity University, Lucknow, India 226010
deepakarorainbox@gmail.com

Abstract:
Today, web has become a most popular trend in terms of availability of rich contents related to almost every field of life. Nowadays, it is emerged as a most demanding tool for searching and retrieving information over a large repository of web contents. Content or resource searching always has been very important for scientists and research scholars. Today, market is full of different search tools over web having noticeable diversity in terms of functioning and the end search results. Given a query, search tools normally return a large number of relevant web pages. To be more effective, the returned pages must be ranked according to their relevancy with respect to the user’s query. Page Rank and Weighted Page Rank Algorithms give the efficient result but these algorithms are query independent algorithms as these are based on the web structure mining. In order to give more efficient result, this paper presents a new algorithm which considers web structure mining and web content mining towards ranking of web pages in accordance to the relevancy of the user’s query.

Keywords: Search Engine; Web Mining; Page Rank; Weight Link.

1. Introduction
The World Wide Web [4] is the assemblage of many interlinked hypertext documents which are accessed via the Internet. Search Engine is the one which helps user to search the documents in this large repository as per the user’s query. These searched documents can be in hundreds which leads to the problem of selecting which document should appear first, second and so on. In order to select this, the concept of Web Page Ranking is used which accomplishes the task of providing a rank to every web page through an algorithm and the web page with higher rank will appear first. Search Engine [7] has a process that goes from Crawling, Indexing, Searching and Ranking of information.

A crawler or spider is a program that seeks websites and scans the contents and other information of their pages in order to generate copy of the visited pages for a search engine index to provide fast searches. This process is called crawling for providing new or updated data that has been submitted by website owners. Indexer is a program that extracts the terms from web pages and creates an alphabetical order of terms. It also contains extra information such as URL of the page, frequency and position of the terms. It is also called as “Search Engine Database”.

Search engine searches the web page in the index created by the indexer in response to the user query and aligns the web page links as per their page rank. The page link with higher page rank will be appeared earlier than the page link with lower page rank. Page Rank is a numeric value representing how important a page is on the web.
There are various ranking algorithms. In this paper, we will discuss only the Page Rank algorithm and the proposed improvement of page Rank. The proposed modified algorithm is based on the web structure mining and the web content mining. It specifies the structure of the web pages and the content of the pages which gives the most relevant web pages.

![Simplified Search Engine Architecture](image)

**Fig. 1:** Simplified Search Engine Architecture

Web Mining [6] is the data mining technique that is used to discover the content of the web, link structure of the web pages and the user’s behavior in the past. Web mining can be categorized as Web Structure Mining (WSM), Web Content Mining (WCM) and Web Usage Mining (WUM). WSM is the process of finding out the relationship between web pages by analyzing web structure or web graph. Web graph consists of web pages as nodes and hyperlinks as edges connecting two pages. WCM is responsible for extracting the relevant or useful information from the content of the web pages. WUM identifies user profile and its behavior as recorded in the web log file.

2. **Background and Related Work**

In the present scenario, content or resource searching is most widely used on WWW. The amount of information on the web is increasing day by day that generates the biggest challenge for information retrieval. In order to fulfill the requirements of user and due to the increasing size of the web, there is a need to perform efficient searching. Ranking the web pages is also a major part for the searched web pages as it is not possible for the user to view all the returned pages. There are various algorithms for ranking the web pages. Among them Page Rank proposed by Brin and Page [1, 2] in 1998 and Hypertext Induced Topic Selection (HITS) algorithm proposed by Kleinberg [5] in 1998 are used successfully in the area of web structure mining. An extension to Page Rank is Weighted Page Rank algorithm proposed by Wenpu Xing and Ali Ghorbani in 2004. This algorithm [10] assigns larger rank value to more important outlinked page while Page Rank divides the rank value of a page evenly among its outlinked pages. The importance of the outlinked page can be measured by its number of inlinks and outlinks. A new ranking algorithm based on the link and visits of the web pages was proposed by A. K. Sharma and Neelam Duhan [8] in 2010. This is based on the web structure mining and web usage mining to take the user visits into account to determine the relevancy of the web pages. Hema Dubey and Prof. B. N. Roy [3] proposed a normalized web page ranking algorithm based on the mean value of the page ranks. This reduces the time complexity of the traditional page rank algorithm by reducing the number of iterations.

3. **Traditional Page Rank Algorithm**

Page Rank is used to determine the importance of the page on the web. Surjey Brin and Larry Page [1] [3] proposed a ranking algorithm named PageRank (PR) that uses the link structure of the web to determine the importance of web pages. According to this algorithm, if a page has important links to it then its links to other pages also become important. Therefore, it takes back links into account and propagates the ranking through links. In Page Rank, the rank score of a page is equally divided among its outgoing links and that values of outgoing links are in turn used to calculate the ranks of pages pointed by that page.
Page A is a backlink of page B and page C while page B and page C are backlinks of page D.

The Page Rank algorithm is given by :
1) Initially, let PAGE RANK of all web pages to be 1.
2) Calculate page ranks of all pages by following formula:
\[
PR(P) = (1-d) + d \left( \frac{PR(T_i)}{C(T_i)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)
\]

Where
- \(PR(P)\) is the Page Rank of page \(P\),
- \(PR(T_i)\) is the Page Rank of pages \(T_i\) which link to page \(P\),
- \(C(T_i)\) is the number of outbound links on page \(T_i\) and
- \(d\) is a damping factor which can be set between 0 and 1, but it is usually set to 0.5
3) Repeat step 2 until values of two consecutive iterations match.

Here, \(d\) is used as the probability of the links followed by the users and \((1-d)\) as the Page Rank distribution from non-directly linked pages.

In this algorithm, the rank score of a web page is divided evenly over its outlinked pages. Pages that are not relevant to the user query may get the higher rank. To overcome this problem, we are proposing a new query dependent algorithm which is based on the web structure mining and web content mining.

4. Proposed Page Rank Algorithm

Although, there are various algorithms that are used by many search engines but the user may not get the required relevant documents on the top pages. To resolve this problem, we are proposing a new algorithm which employs web structure mining and web content mining together in order to get the relevant web document. Web structure mining is used to get the linking structure of the web pages and web content mining is used for the content of the web pages. Through web content mining, relevant web pages can be fetched as per the user’s query.

**Weight of outbound link:**
According to Wenpu Xing and Ali Ghorbani’s “Weighted Page Rank Algorithm”, \(WL(v, u)\) is the weight of each outbound link connecting pages \(u\) and \(v\) calculated by :
\[
WL(v, u) = \frac{I(u)}{\sum_{p \in R(v)} I(p)}
\]

Where \(I(u)\) and \(I(p)\) represent the number of incoming links of page \(u\) and page \(p\) respectively. \(R(v)\) represents the reference page list of page \(v\).
The content based page rank algorithm is given as:
1) Initially, let PAGE RANK of all web pages to be 1.
2) Calculate page ranks of all pages by following formula:
   \[ PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v)WL(v, u)W_c \]  ..............(2)
   
   Where
   \( PR(u) \) and \( PR(v) \) are the Page Rank scores of page \( u \) and \( v \) respectively,
   \( B(u) \) is the set of pages that point to \( u \),
   \( W_c \) is the content weight \([9]\) of the web pages with respect to the query terms. It is the ratio of the sum of the frequency of the possible query strings \((J)\) and sum of the frequency of whole query string \((Y)\).
3) Repeat step 2 until values of two consecutive iterations match.

5. Comparison of the Proposed Algorithm with the Traditional Algorithm

We will consider an example of hyperlinked structure as shown in Figure.3 in order to explain the working of page rank. Here, A, B and C are three web pages. Iterative method is used for calculating the page rank. Each page is assigned a starting page rank value of 1.

By the Traditional Page Rank algorithm, the page ranks of the pages A, B and C can be calculated using equation (1) as shown below –

\[
PR(A) = (1-d) + d(PR(B)/2 + PR(C)/2)
\]
\[
PR(B) = (1-d) + d(PR(C)/2)
\]
\[
PR(C) = (1-d) + d(PR(A)/1 + PR(B)/2)
\]

![Fig.3: An Example of hyperlinked Structure](image)

By Calculating these equations with \( d=0.5 \) (say), the page ranks of the three pages are:
\( PR(A) = 1.2, PR(B) = 0.8, PR(C) = 1.2 \)
Here, \( PR(A) = PR(C) > PR(B) \)

By the proposed content based page rank algorithm, the equations can be calculated using equation (2) as shown below :

\[
PR(A) = (1-d) + d(\{PR(B)WL(B, A)W_c\} + \{PR(C)WL(C, A)W_c\})
\]
\[
PR(B) = (1-d) + d(PR(C)WL(C, B)W_c)
\]
\[
PR(C) = (1-d) + d(\{PR(A)WL(A, C)W_c\} + \{PR(B)WL(B, C)W_c\})
\]

Now, the WeightLinks can be calculated as:
\[
WL(B, A) = \frac{I(A)}{\sum_{p \in R(B)} I(p)}
\]
\[
WL(B, A) = \frac{I(A)}{I(A) + I(C)} = \frac{2}{2+2} = \frac{1}{2}
\]
Similarly,

\[ WL(C, A) = \frac{2}{3}, WL(C, B) = \frac{1}{3} \]

And

\[ WL(A, C) = 1, WL(B, C) = \frac{1}{2} \]

Suppose, Content Weights of the three pages A, B and C are 10/12, 11/17 and 7/15 respectively. Now, After Substituting \( d = 0.5 \), content weights and above calculated weight values, the page rank obtained as :

\[ PR(A) = 0.72, PR(B) = 0.57, PR(C) = 0.90 \]

It can be seen here that

\[ PR(C) > PR(A) > PR(B) \]

It shows that the resulting order of the pages obtained by PageRank and the proposed mechanism are different. In PageRank, Pages A and C have the same importance but in the proposed mechanism Page C has more importance than the Page A.

6. Conclusion

Generally, users spend a lot of time in search of the desired information. The paper presented a modified page rank mechanism that provides the important and relevant results than the original page rank algorithm as the modified one uses web structure mining as well as web content mining. In Traditional web PageRank algorithm, the rank score of a web page is divided evenly over its outlinks while the proposed PageRank algorithm assigns larger rank values to more important and more relevant outlinks. Traditional PageRank algorithm is based on the link structure of the web pages while proposed PageRank algorithm is based on the link structure and contents of the web pages. As a result, user may find the desired content in the top few pages. Thus, the search space can be reduced at some extent.

Acknowledgments

We are very thankful to respected Mr. Aseem Chauhan, Chairman, Amity University, Lucknow, Maj. Gen. K. K. Ohri, AVSM (Retd.), Pro-Vice Chancellor, Amity University, Lucknow, India, for providing excellent computation facilities in the University campus. We also pay our regards to Prof. S.T.H. Abidi, Director and Brig. U.K. Chopra, Deputy Director, Amity School of Engineering and Technology, Amity University, Lucknow for giving their moral support and help to carry out this research work.

References