A Hybrid Page Rank Algorithm for Web Pages
Abhishek Sharma$^{1}$ Arun Solanki$^{2}$

1,2School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India

Abstract— In the past, the World Wide Web has taken a rapid growth and search engines are the most popular ways of finding any kind of information on it. There are many cases, where the users are having so many web pages where they need to find there solutions in response to the searched query. Generally there are not many users that go beyond the first few pages. Hence the search engines face a challenge on how to provide the best meaningful web pages in response for a user query. In spite of having so many page rank algorithms used by the search engines, the pages that are needed by the users may actually get lost in the tremendous amount of information returned by the search engines. The users that use the search engines are not expert in the domain of search engines, hence the information need to be grouped that is provided by the web pages so that they can navigate the information quickly. Hence the analysis of the pages importance would be done by various matrices and enable the representation of interactions between pages in the form of a network. And finally the analysis of what is the efficiency of rank values of the pages using the proposed system and existing system is carried out. For this purpose, algorithms from the discipline of graph theory are applied, including the Dijkstra’s algorithm for impact analysis, Page Rank and Weighted Page Rank algorithm for calculating the rank values of each web page are used. The weighted page rank algorithm is used to find the page rank of the web pages and with the help of Dijkstra’s algorithm an optimal path is decided on the basis of in-links and out-links. The Dijkstra’s algorithm gives the direction to the Weighted Page Rank algorithm to calculate the page rank values of the pages. This research proposed a Hybrid Page Rank System (HPRS). The HPRS system is composed of three modules User Interface, Operation module and Knowledge Base. The tools used are visual basic server and SQL server database and the framework on which the system is implemented is dot net. The proposed system is implemented in asp.net language. The result of the research compares the rank values of the existing and the proposed system. The performance is calculated for each page simultaneously, is done between the existing Page Rank Algorithm and the Hybrid Page Rank Algorithm. The proposed system gives more efficient rank values of the pages than the existing algorithm.

Key words: Page Rank Algorithm, Weighted Page Rank Algorithm, Dijkstra etc

I. INTRODUCTION

The Internet is a global system of interconnected computer networks that use the standard Internet protocol suite (TCP/IP) to link several billion devices worldwide. It is defined as a network of networks that consists of millions of private, public, academic, business, and government networks of local to global scope, linked by a broad array of electronic, wireless, and optical networking technologies. The Internet carries an extensive range of information resources and services, such as the inter-linked hypertext documents and applications of the World Wide Web (WWW), the infrastructure to support email, and peer-to-peer networks for file sharing and telephony. WWW is an information system that is accessed via the internet. It is a system of interlinked hypertext documents. It is also known as web. The individual web pages on the internet are known as web pages and are accessed by various browsers used in the computer system. These pages consist of text, videos and other multimedia components as well as hyperlinks. The web pages are search and ranked on the bases of an engine known as search engine. Search engines are the programs that search the documents for specified keywords and returns a list of the documents for the keywords found. A general class of programs is a search engine, however, this term is often used to specifically describe systems like Google, Bing and Yahoo! It is a Search that enables users to search for documents on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). There are basically two different types of search engines:

1) Crawler Based Search Engine: Crawler based search engine includes Google, AllTheWeb and AltaVista etc.

2) Human-powered directories: Human powered directories include Yahoo directory, Open Directory and LookSmart etc.

The PR algorithm is a recursive algorithm in which the rank of a page is influenced by the number of in links and the page rank of the pages linking to it. Therefore a page confers its importance to the pages by evenly distributing its page rank value to all its outlinks whereas WPR algorithm assigns a larger page rank values to the most important pages rather than dividing the rank value evenly among all its outgoing links. Therefore each outgoing link gets a value proportion to its importance and the importance is assigned in terms of weight values to the incoming and the outgoing links. Firstly, the optimised path to a page is found by the dijkstra’s algorithm and then we calculate the page rank using both page rank(PR) and Weighted Page Rank (WPR) algorithm. Finally the comparison is done between PR and WPR. This research has used dijkstra’s algorithm instead of warshall’s algorithm as the time complexity of dijkstra’s algorithm is O(n^2) which is better than Warshall’s i.e. O(n^4).

II. RELATED WORK

The related work of the algorithm is as follows:

According to the author H. Dubey, the PR algorithm is discussed which shows to calculate the importance of the pages using the same algorithm.

According to the author (T.Kumari et al.), the formula for calculating the importance of the pages for PR and WPR algorithm is taken. These papers also tell that WPR algorithm is better than the PR algorithm. According to the authors (F.Mokammel, et al., 2012) in the report Methodology of Quality Grade and Change Impact Analysis.
of Requirements. They have analyze the requirement phase for its quality by using the integrated requirements matrices and convert them into graphical structure by the equation. According to the authors (D.J. Highamy and A.Taylorz, 2003) in the research paper “The Sleakest Link Algorithm” have given the criteria of how to calculate PR of the web pages using Jacobi Iterator. According to the author (S.Dominich, 2005), in the report “Page Rank And Interaction Information Retrieval” have given the criteria to calculate the page rank of web pages using mathematical models. According to (H.Dubey and Prof. B. N. Roy, 2011) in the research paper An Improved Page Rank Algorithm based on Optimized Normalization Technique learned the improved page rank algorithm and its working. A new page rank algorithm which uses a normalization technique based on mean value of page ranks. According to (D.K.Sharma and A.K.Sharma,2010) in the research paper A Comparative Analysis of Web Page Ranking Algorithms the calculation of the page rank of the inbound and outbound pages is determined. This paper deals with analysis and comparison of web page ranking algorithms based on various parameter to find out their advantages and limitations for the ranking of the web pages. According to (S.Hougardy,2010) in the research paper The Floyd-Warshall Algorithm on Graphs with Negative Cycles which tells that Floyd-Warshall algorithm exponentially large numbers may occur, if the input graph contains negative cycles. This shows that for larger graphs it can be quite likely to have sub graph causing an overflow and thus can have high time complexity. According to (K.Magzhan and H.M.Jani, 2013) in the research paper A Review And Evaluations Of Shortest Path Algorithms the Warshall’s algorithm has high complexity than Dijkstra’s algorithm. According to (B.M.Maggs and S.A.Plotkin,2002) in the research paper Minimum-cost spanning tree as a path-finding problem Warshall algorithm has high time complexity than the same in Dijkstra’s algorithm.

III. PROPOSED ALGORITHM

This work includes developing website on .NET 2010 framework and generating tables in SQL SERVER 2005. An interface is developed on .NET 2010 framework for calculating page rank and weighted page rank for the pages. Assigning of weight to each of the web pages for sequential access pattern matching.

This generates the graph between the Page Rank, Weighted Page Rank Algorithm and Hybrid Algorithm in Sequential Pattern.

There are the equations used in this implementation.

D : array of distances from the source to each vertex
prev : array of pointers to preceding vertices
i : loop counter index
F : list of finished vertices
U : unfinished vertices

/* Initialization: set every distance to (total number of inlink/outlink) until we cover previous position */

Calculate Win (v,u): Then calculate the Win (v,u) for each node present in web graph by applying the equation (1) as below.

The popularity from the number of inlinks and outlinks is recorded as Win(v,u) and Wout(v,u), respectively.

\[ \text{Win} = \frac{I_u}{p} \quad \text{equation (1)} \]

where Iu and Ip represent the number of inlinks of page u and page p, respectively. R(v) denotes the reference page list of page v.

Similarly Calculate Wout (v,u): Then calculate the Wout (v,u) for each node present in web graph by applying the equation (2) as below.

\[ \text{Wout} = \frac{O_u}{O_p} \quad \text{equation (2)} \]

for i = 0 to |V| - 1

D[i] = Win/Wout

prev[i] = NULL

end

/* The distance from the source to the source is defined to be zero */
D[s] = 0

/* This loop corresponds to sending out the explorers walking the paths, where
while(F is missing a vertex)
pick the vertex, v, in U with the shortest path to s
add v to F
for each edge of v, (v1, v2)
  /* The next step is sometimes given the confusing
  name "relaxation"
  if(D[v1] + length(v1, v2) < D[v2])
    \[ D[v2] = D[v1] + length(v1, v2) \]
    prev[v2] = p1
Page A has two reference pages: p1 and p2. Inlinks and outlinks of these two pages are Ip1 = 2, Ip2 = 1, Op1 = 2, and Op2 = 3. Therefore,

\[ W_{in}(A,p1) = \frac{I_p}{I_p + I_p} = 2/3 \quad \text{--- equation (3)} \]

\[ W_{out}(A,p1) = \frac{O_p}{O_p + O_p} = 2/5 \quad \text{--- equation (4)} \]

Considering the importance of pages, the original PageRank formula is modified as

\[ \text{PR}(u) = (1 - d) + d \sum_{v \in V} \text{PR}(v) \frac{\text{Win}(v,u) \cdot \text{Wout}(v,u)} {\text{Wout}(v,u)} \quad \text{Equation (5)} \]

possibly update UPR(u) depending on implementation
end if
end for
end while

IV. SYSTEM ARCHITECTURE

A. Architecture of Hybrid Page Rank System (HPRS):
The 3-tier architecture is shown in figure 1:

- This module shows how the system looks to the user. The algorithm consider the quality in-links, out-links and of each page and based on that an adjacency metric is calculated which shows the traceability of each page with all other pages.
- The metric will then used to calculate the overall reachability of the pages using the Dijkstra’s algorithm.
- The resulting metric is used to calculate the rank of each page.
- The results are given back to the user interface.
- The 3-tier architecture is given as:

The 3-tier architecture is divided into 3 modules namely
A Hybrid Page Rank Algorithm for Web Pages

(IJSRD/Vol. 3/Issue 0/2015/918)

All rights reserved by www.ijsrd.com

Fig. 1: 3-tier Architecture

The module which contains the procedures/methods/generators are used to calculate the map the requirements in a set of values adjacency metrics. It includes the following sub modules as shown in Figure 1.

1) **Operation Module**

This module is further divided into the following sub modules which are explained below:

a) **Traceability Generator**
   The module generates the traceability metrics which is also known as adjacency metric. This metric is generated after comparing each requirement.

b) **Page Rank Metric Generator**
   This module calculates the importance of the pages using the existing PR algorithm. The rank of the pages are decided when the two iterations in the traceability metric are nearly equal.

c) **Hprs Metric Generator**
   The module will calculate the rank of the each page using the hybridization of WPR and Dijkstra’s algorithm which tells their importance level.

2) **Knowledge Base**

The knowledge base is the backend database which is used to store the requirements in words as well their integers, float and may be double values. The knowledge management and acquisition module manages the data stored in the database. The knowledge base is of two types static and dynamic.

3) **User Interface**

The interface where user can see the working of the proposed algorithm i.e. the page rank of the pages using existing PR algorithm, the page rank of the pages using proposed algorithm and their comparison.

a) **Website GUI**
   This interface captures the in-links and out-links of an offline web application and stores the captured values in SQL server database.

b) **Inlinks and Outlinks GUI**:
   The captured in-links and out-links for each page are shown in the GUI dynamically.

c) **Rank Metric GUI**
   This GUI calculates the rank of each page from the in-links and out-links values captured and maps them to the adjacency metrics.

d) **Comparison GUI**
   This GUI compares the PR and HPR and the result is shown in the form of a graph.

B. **Process Flow of Hybrid Page Rank System**

The process flow diagram is shown in figure 2.

1) **STEP 1:**
   The user interacts with the GUI by selecting the pages of the offline web application. A website GUI is made which is offline and some inlinks and outlinks are mentioned in it.

2) **STEP 2:**
   The system GUI is built using Microsoft Visual Studio framework and consist of the website GUI and hybrid page rank system GUI.

3) **STEP 3:**
   Through the system GUI the count of the in-links and out-links of the pages are gathered from the offline web applications and gets stored in sql server. The values stored in the database are dynamic values. For in-links and out-links separate tables are made.

4) **STEP 4:**
   The traceability metric is calculated from the captured requirements in step3 which consist of many iterations. These iteration continues till the time when the two iterations continuously nearly similar.

5) **STEP 5:**
   The final iteration are captured in the list. These values are used to determine which page has maximum page rank.

6) **STEP 6:**
   The importance of the pages is calculated using PR algorithm and the page with highest page rank value is known. The page with maximum page rank is displayed on the GUI.

7) **STEP 7:**
   The importance of the pages are calculated
using HPR algorithm.

8) STEP 8: The page rank of the pages is identified. The importance of the pages calculated using HPR algorithm has less value as compared to the PR algorithm. The comparison graph is displayed to the user interface.

V. WORKING AND IMPLEMENTATION OF HYBRID PAGE RANK SYSTEM

A. Step 1: Storing Values in Knowledge Base

B. Step 2: Getting URLs

C. Step 3: Getting the Inlinks and Outlinks

The links and o gathered are stored in the database. The database consists of both the tables of inlinks and out-links as shown in figure 3.

B. Step 2: Getting URLs

The following table is listing URLs of the off line websites which have inlinks and out-links as well as shown in figure 4.

These are the pages of the offline application, each of which has in-links and out-links to each other. From this list, one page is selected to get the inlinks and outlinks of the same. The inlinks and outlinks comes dynamically from the database.
from the database, it these values are shown in the user interface as shown in figure 5.

D. Step 4: Rank Metric Formation

![Rank Metric Formation](image1.png)

The iterations constitutes the rank metric or traceability metric, the iterations are carried till the time there are almost two similar iterations, final iterations that comes out gives the rank of the pages as shown in figure 6.

E. Step5: Page Rank Evaluation

The results produced from the final iterations gives the page rank using the proposed system and the values are sent to the user GUI where it is displayed shown in figure 7.

![Page Rank Evaluation](image2.png)

F. Step 6: Applying Hybrid Page Rank(HPR)

![Hybrid Page Rank Algorithm](image3.png)

As shown in figure 8, the rank values using the proposed algorithm are carried out. HPR algorithm gives the rank values of each page. The ranks values are 1.11, 0.69, 0.68, 1.10, 0.51.

G. Step 7: Comparison

![Comparison](image4.png)

After the evaluation of weighted page rank, a comparison is done between the existing and proposed page rank algorithm. The graph is displayed by clicking the button comparison graph which is shown in figure 9. The comparison will be shown in Crystal Viewer(CV).

H. Step 8: Graph Generation

![Performance Comparison](image5.png)

The performance on the basics of the rank values for both the existing algorithm and proposed algorithm is shown in figure 10.

VI. Result

The rank values for each page using the Page Rank and Hybrid Page Rank algorithm is taken and a comparison is done in each of them for each page simultaneously in table 1.

A. Comparison between PR and HPRS

<table>
<thead>
<tr>
<th>PAGES</th>
<th>Existing (PR)</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://localhost:1232/linking/Default.aspx">http://localhost:1232/linking/Default.aspx</a></td>
<td>1.27</td>
<td>1.11</td>
</tr>
</tbody>
</table>
Table 5.1: Comparison between PR and Hybrid Page Rank
The table shows that the rank values using existing algorithm are 1.27, 1.24, 0.98, 1.22 and 0.66 and the same using the Hybrid Page Rank algorithm are 1.11, 0.69, 0.68, 1.10 and 0.51. Therefore it is proved that the HPRS algorithm gives higher efficient rank values to the pages than the PR algorithm. Hence, it is proved that the proposed algorithm is more efficient than the PR algorithm.

B. Discussion
1) Rank Values Comparison between Existing And Hybrid Page Rank:
The ranks of web pages using the Page rank and Hybrid page rank algorithm are compared. The plot of the table is given by the bar graph and is shown by the figure 5.1. In the horizontal axis, there are number of pages that are used in the system and on the vertical axis, the rank values are shown.

2) Performance Comparison between PR and HPR Algorithm:
The performance is calculated in terms of rank values of the web pages for PR and HPR. The formula for determining the performance of page say “about us” is defined below by equation 6.

\[
\text{Performance(About Us) = } \frac{\text{Page Rank(About Us)}}{\sum_{n \in \text{Page Rank(n)}}}
\]

Equation (6)

Similarly, the performance of each page can be calculated. The performance while calculating the rank values for both the Page Rank and Hybrid Page Rank is shown in graph 2.

VII. CONCLUSION
Traditionally, the ranking of the web pages are done using the page rank algorithm, weighted page rank algorithm and many other. The in-links and out-links of the web pages are taken into considerations and using them the page rank values are provided to each web page. The Hybrid Page Rank algorithm calculates the page rank of the pages by choosing the optimal path between the web pages which is provided by Dijkstra’s algorithm. The dijkstra’s algorithm in the hybrid page rank algorithm tells which path to choose when we have same weights of two pages. It also helps to consider the weights of the indirect pages.

The Hybrid Page Rank System (HPRS) is composed of three modules A) User Interface B) Operation Module C) Knowledge Base. The user interface provide the Website Graphic User Interface (GUI) in which the web pages has in-links and out-links and HPRS GUI which further has three GUI 1) the in-links and out-links GUI, where the in-links and out-links of the web pages are displayed. 2) the Rank metric GUI, where the page rank of the web pages are displayed. 3) Comparison GUI, where the comparison of proposed and existing algorithm are compared. The implementation of the proposed system is illustrated in chapter 4 and the results obtained for the proposed system is illustrated in chapter 5.

The results obtained of the proposed system are compared with the existing approach taken from the literature. The result shows that the HPRS algorithm gives more efficient rank values to the pages than the existing algorithm therefore it is proved that the proposed algorithm is more optimized than the existing algorithm. Hybrid Page Rank Algorithm are implemented and the ranking and traversing of web pages are improved.

In the proposed algorithm weight is assigned to each of the web pages as on in-links or out-links pages traversing. The proposed algorithm is efficient in finding of the web pages importance based on the ranking.

A. Suggestion for Further Research
On the basis of studies and investigations carried out in this research, the following suggestions are presented for future research.

- The proposed method may be extended to work on work on naive byes’ classifier, to check the
maximum probability of the occurrence by enhancing the page rank technique. The current method uses the weighted page rank technique.

- The proposed method may be used for hit count in web page, which may be helpful in text mining approach.
- Future developments of the tool may go towards the development of functionalities supporting the interactive design of requirements models.

The proposed algorithm may be optimizing by using various optimizing techniques. This research not uses any optimizing technique.

REFERENCES


