9. CONCLUSION

The objective of doing research in Web Mining is to determine efficient hyperlink analysis in Web Structure Mining. The next goal of this research is to enhance the performance of Information Retrieval by using the Hybrid Web Search method. This method was designed by incorporating Web Structure Mining (WSM) and Web Content Mining (WCM) for efficient web search. The first objective was obtained by computing PageRank values with non-uniform transition probability distribution using the link structure of the Web pages offline. The next objective was obtained by designing architecture of HWS method, which integrates the result of efficient hyperlink analysis along with Cosine Similarity (CS) score between search query terms and Web pages.

The network structure of a hyperlinked environment can be a rich source of information about the content of the web. Hence web structure mining plays a vital role in web search. Web Structure Mining algorithms explore the hyperlink structure of the web. Therefore, this research work concentrated on four influential hyperlink based search algorithms in WSM. They are

- Hypertext Induced Topic Search (HITS)
- Weighted PageRank Algorithm (WPR)
- PageRank Algorithm (PR)
- Improved PageRank Algorithm (IPR)

HITS is search query dependent. HITS algorithm ranks the web pages by analyzing their in-links and out-links. When the user issues a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings of the expanded set of pages, authority ranking and hub ranking. An
authority is a page with many in-links and the page may have authoritative content on
some topics, hence many users link to it. A hub is a page with many out-links. The
page serves as an organizer of the information on a particular topic and points to
many good authority pages on the topic. The HITS algorithm is evaluated by using
the Clever search engine.

WPR algorithm is an extended PageRank Algorithm, which assigns larger
rank values to more popular pages and does not divide the rank of a page uniformly
among its out-link pages. Each out-link page gets a value proportional to its
popularity. The popularity of a page is determined by observing the number of in-
links and out-links. The WPR algorithm is evaluated by using research model.

PageRank algorithm is a static ranking of web pages which is computed for
each page off-line and it does not depend on search queries. A hyperlink from a page
pointing to another page is an implicit conveyance of authority to the target page. A
page which receives more in-links has more prestige. PageRank score of page $i$ is
determined by summing up the PageRank scores of all pages that point to the page $i$.
Since a page may point to many other pages, and its prestige score should be shared
equally among those pages.

The Page Rank algorithm has the \textit{ranksink} problem. Due to \textit{ranksink}, pages in
the loop to have higher PageRank values than the existence. To solve the \textit{ranksink}
problem a small probability $d/N$ (where $d$ is dampening factor and $N$ is the number of
pages) is added to each entry of the Transition Probability Distribution (TPD) matrix.
Major problem of PageRank algorithm is \textit{dangling pages}. If a page has no outbound
link, then the page is called as a \textit{dangling page}. Due to \textit{dangling page}, PageRank loss
occurs continuously during the iteration process. This event is called as \textit{norm - leak
phenomenon}.
The main characteristic of Improved PageRank algorithm is to overcome the dangling page problem by adding a complete set of outbound links from the *dangling page* to every page including itself with the uniform probability distribution $1/N$. In the Improved PageRank algorithm, PageRank scores might converge to a fixed point and the sum of all PageRank values are always maintained to be one. The PR and IPR algorithms are evaluated by using Google search engine.

The experimental setup of these algorithms is evaluated in JAVA using the workstation (Intel 2nd Generation Core i3 Processor (3.10 GHz) machine with 8GB RAM). Convergence of the page rank values depends upon the volume of the data sets. Dampening factor ‘$d$’ varied between 0 to 1, analyses is carried out for the value of 0.15, 0.5, and 0.85.

Preliminary experiment was conducted with the different sample web graphs to analyze the working functionality of WSM algorithms and to determine the efficient hyperlink analysis.

Further analysis of this study is focused on comparative analysis of PR and IPR algorithms with two different characteristics (i.e *ranksink, dangling*). Using these characteristics, four different case studies were considered. For this analysis, the sample web graph of ten web pages and the dampening factor $d = 0.5$ was considered. The four different case studies are considered as follows:

*Case Study 1: No ranksink and no dangling pages*

*Case Study 2: Has ranksink and no dangling pages*

*Case Study 3: No ranksink and has dangling pages*

*Case Study 4: Has ranksink and has dangling pages*
From the analysis of these case studies, it was observed that the PR and IPR algorithms work identically for case study 1 and 2.

From the analysis of case study 3 and 4 it was observed that the PageRank value generated by the PR algorithm was entirely different from the IPR algorithm. The PR algorithm was executed for 50 iterations and it was observed that the $\text{Rank}_i$ and $\text{Rank}_{i+1}$ never get converge to the same PageRank value, because PageRank loss had occurred. IPR algorithm was executed for 50 iterations, and it was observed that $\text{Rank}_i$ and $\text{Rank}_{i+1}$ get converge at the earlier stage of the iterations itself, which then remains unchanged. From the comparative analysis between these two algorithms, it was observed that the IPR algorithm produces better ranks.

Two matrices namely Initial Probability Distribution (IPD) matrix and Transition Probability Distribution (TPD) matrix were used in IPR algorithm to compute the rank of web pages and the entries of these two matrices were uniformly distributed.

To obtain the objective of the efficient hyperlink analysis, suitable modification was carried out in the input of IPR algorithm, which leads to Proportionate Prestige Score (PPS) method. In proposed PPS method the prestige of each page is proportionally distributed according to the prestige of the page to which it links. The prestige of a web page is more important to determine the rank of a web page, which could be computed exactly by using PPS method. This proposed PPS method computes the PageRank values using non-uniform transition probability distribution of prestige values of each page, which is applicable to Initial Probability Distribution (IPD) Matrix and Transition Probability Distribution (TPD) Matrix.

Initial probability distribution vector for each page is measured based on the inbound links of a web page. The entries of IPD matrix for each page is taken as $\frac{a_i}{g}$. 
where \( \alpha \) denotes inbound link of each page and \( \beta \) denotes the total number of links of web graph. The sum of IPD matrix entries should be one and these entries are in the interval of 0 to 1.

Using the proportional distribution of prestige values, the entries of TPD matrix are assigned non-uniform transition probability distribution values. The sum of each column of TPD matrix should be one and these entries are in the interval of 0 to 1. For the dangling pages the weight of out-links of dangling page is computed according to the prestige of all web pages including itself and it is proportionally distributed.

The following steps are used to compute the entries of TPD matrix

1. Prestige of each web page can be calculated by adding the weight of inbound links of each web page
2. If a web page has many outbound links to different pages, then summing the prestige of those pages is considered as Total Prestige.
3. Each element of TPD Matrix is obtained by dividing the prestige of each page by its corresponding Total Prestige values.

Working functionality of the PPS method has been demonstrated with a sample web graph consisting of 20 pages and it was compared with IPR algorithm. From this analysis, PPS Method gives less weightage to the irrelevant web pages and more weightage to the most relevant web pages.

Further study of this research work, the working functionality of IPR and PPS were tested with huge number of web pages from 300 to 4000 with randomly chosen in-links and out-links. From the analysis, this study proves that the proposed PPS method boosts up the most significant pages on an average of 44% and pulls down the least significant pages on an average of 32% when compared with the IPR algorithm.
To substantiate the efficiency of PPS method, further analysis has been focused on testing the efficiency of the proposed PPS method and comparing it with the IPR algorithm using benchmarked on real social network data from three different domains: Social Circle: Facebook, Wikipedia vote network and Enron email network. All datasets are collected from Stanford Large Dataset Collection repository. The maximum number of iterations and the dampening factor for computation of rank values are set as 50 and 0.85, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Web Pages</th>
<th>Number of Hyperlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset I : Social Circle: Facebook</strong></td>
<td>4039</td>
<td>88234</td>
</tr>
<tr>
<td><strong>Dataset II : Wikipedia vote network</strong></td>
<td>7115</td>
<td>103689</td>
</tr>
<tr>
<td><strong>Dataset III : Enron Email Network</strong></td>
<td>36692</td>
<td>367662</td>
</tr>
</tbody>
</table>

From these analyses, it was observed that the order of the search result was improved while using the PPS method. Rank score of most significant pages were boosted up by 14% and irrelevant and weak relevant pages were pulled down by 0.14% for the data set III. Spearman's Rank correlation was applied in PPS to determine the correlation between PPS and IPR algorithm, which produce positive correlation between them.

In addition to the WSM algorithms, this research work also focused on WCM techniques, which is also an important factor for ranking the web pages. Therefore, this research work is considered two existing relevancy score methods in WCM namely

- Three Level Scoring (TLS) Method.
- Vector Space Model (VSM)
The working methodology of TLS method has been demonstrated with content of different sample web graphs. These web graphs consist of 15 and 20 web pages for the query strings “Consulting Service Management” and “Business Application Development” were randomly selected from the web sites http://www.wipro.com and http://www.tcs.com to determine the rank of web pages. From these experiments, it was observed that TLS gives weightage to the pages based on the order of the query terms appear in the web page and preprocessing operations are not performed. Therefore, this research work consider another relevancy score method namely VSM. The working functionality of VSM has been demonstrated with the sample web graph. This web graph consists of 12 web pages in the form of text files for the query string “Knowledge Discovery”, which was randomly selected from the website http://en.wikipedia.org/. In VSM, pre-processing operation was carried out such as stop words were eliminated from the web pages using the list of stop words (http://www.webconfs.com/stop-words) and stemming was performed using the Porter Stemmer algorithm (http://tartarus.org/martin/PorterStemmer). From this experiment, it was observed that VSM is better than TLS, hence VSM is most widely used in the traditional Information Retrieval.

Another objective of this research work focuses on hybrid web search method by incorporating WSM and WCM for effective web search. To obtain this objective, this research work focused on the architecture of the hybrid web search method. Using this architecture the following HWS methods were analyzed with different sample web graphs.

- HWS Method using WPR with TLS - WPR_TLS
- HWS Method using WPR with VSM - WPR_VSM
- HWS Method using IPR with VSM - IPR_VSM
- HWS Method using PPS with VSM - PPS_VSM
Procedure of WPR_TLS has been demonstrated with the sample web graph of twenty web pages for the query string “Business Application Development” From this experiment, it was observed that the WPR_TLS is not producing an improved result when compared with other HWS methods. Therefore, this research considers the procedures of WPR_VSM, IPR_VSM and PPS_VSM. These HWS methods have been illustrated with the sample web graph of 12 web pages for the query string “Knowledge Discovery”. The results of these methods are also compared with each other. From these experiments, it was observed that the PPS_VSM and IPR_VSM produce better results when compared with WPR_VSM. Therefore, further analysis of this research work was concentrated on PPS_VSM and IPR_VSM and compared with each other.

Procedures of PPS_VSM and IPR_VSM were tested with huge number of web pages such as 300, 500, 1000 and 2500 web pages for the query strings "web", "Data Mining", "Web Structure Mining" and “Web Data Mining” respectively. From these analyses, it was identified that the PPS_VSM out performed, when compared with IPR_VSM method. This present study proves that the proposed PPS_VSM boosts up the most significant pages by 34% in an average, and pulls down the least significant pages in an average of 30% when compared with IPR_VSM. These two HWS methods show positive correlation when the Spearman's Rank correlation was applied, which is an expected characteristic in the proposed methods too. The order of the search result is also improved according to the requirement of the user.

As a future scope, this proposed PPS method can be analyzed for its performance in other Web Structure Mining algorithms and also proceeds by using fuzzy based information retrieval system to personalize the link based search results with respect to user’s interests.
FINDINGS

- Irrespective of the entries of IPD matrix, the PPS always produces high score. Hence the entries of the TPD matrix always play a vital role in deciding the rank score of web pages.
- In the computation of rank score for the dampening value equal to 0.85, the proposed PPS method shows an observable improvement.
- Rank scores of the most relevant pages are significantly boosted up and the least relevant pages are pulled down in the search result.
- Using the non-uniform probability distribution of the TPD and IPD matrices, this research work proves that the iteration process of the PageRank calculation always converges to a fixed point.
- Applying non-uniform transition probability distribution of prestige on the web graph, it overcomes the dangling page problem in an efficient way.
- High prestigious page either pointing to or pointed by other pages was more important than the less prestigious page, when the prestige of a web page is proportionally distributed.
- Efficient hyperlink and effective content analysis can also be incorporated in the architecture of Hybrid Web Search method for improving the search results. Significant percentage of improvement was obtained by using HWS method is given below

<table>
<thead>
<tr>
<th>Web Page</th>
<th>IPR_VSM</th>
<th>PPS_VSM</th>
<th>% of Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>Max: 0.005106, Min: 0.00186</td>
<td>Max: 0.007498, Min: 0.001159</td>
<td>46.85%</td>
</tr>
<tr>
<td>500</td>
<td>Max: 0.003352, Min: 0.001221</td>
<td>Max: 0.005153, Min: 0.000786</td>
<td>53.73%</td>
</tr>
<tr>
<td>1000</td>
<td>Max: 0.00153, Min: 0.000766</td>
<td>Max: 0.00193, Min: 0.000593</td>
<td>26.14%</td>
</tr>
<tr>
<td>2500</td>
<td>Max: 0.000868, Min: 0.000304</td>
<td>Max: 0.000962, Min: 0.000234</td>
<td>10.83%</td>
</tr>
</tbody>
</table>
Real datasets are used to compare the performance of PPS and IPR algorithm. A significant percentage of relevant web pages revealed on the top most of web search results by using the PPS method. The results of the real datasets used for comparative analysis of PPS and IPR algorithm is given below.

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>IPR Max</th>
<th>IPR Min</th>
<th>PPS Max</th>
<th>PPS Min</th>
<th>% of Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset I</td>
<td>0.000767475</td>
<td>0.000214146</td>
<td>0.00178564</td>
<td>0.000210808</td>
<td>14.5%</td>
</tr>
<tr>
<td>Dataset II</td>
<td>0.001246791</td>
<td>0.000107317</td>
<td>0.003517063</td>
<td>0.000103706</td>
<td>182.1%</td>
</tr>
<tr>
<td>Dataset III</td>
<td>0.0044076578</td>
<td>0.0000231919</td>
<td>0.0050469949</td>
<td>2.31588E-05</td>
<td>132.7%</td>
</tr>
</tbody>
</table>

Efficient web search can be incorporated by using this research innovation.