Application of Page Ranking Algorithm in Web Mining

A.M. Sote\textsuperscript{1}, Dr. S. R. Pande\textsuperscript{2}

\textsuperscript{1}Lecturer, Department of Computer Science, ACS College, Arvi, India, \\
\textsuperscript{2}Associate Professor and Head, Department of Computer Science, SSESA’s Science College, Nagpur, India

\textbf{ABSTRACT}: The World Wide Web is a popular and interactive medium to disseminate information today. It is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks. It is very difficult for a user to find the high quality information which he wants to need. When we search any information on the web, the number of URL’s has been opened. User wants to show the relevant on the top of the list. So that Page Ranking algorithm is needed which provide the higher ranking to the important pages. In this paper, we discuss the Page Ranking algorithm to provide the higher ranking to important pages.

\textbf{Keywords} - HITS, Page Rank Algorithm, Web mining, Weighted page Rank,

\section{INTRODUCTION}

The World Wide Web is a very useful and interactive resource of information like hypertext, multimedia etc. When we search any information on the Google, there are many URL’s has been opened. The bulk amount of information becomes very difficult for the users to find, extract and filter the relevant information. So that web mining techniques are used to solve these problems.

Web mining is the application of Data Mining technique to find useful information from web data. With the help of web, we can access multiple data. In the distributed information environment, document or objects are usually linked together to facilitate interactive access to that we can easily access information. There are some following tasks: \cite{1}

1. \textit{Resource finding}: It is the process which involves extracting data from either online or offline text resource available on the web.
2. \textit{Information selection and pre-processing}: The automatic selection and pre-processing of particular information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in the training corpus and representing the text in the first order logic form.
3. \textit{Generalization}: It automatically discovers specific patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization.
4. \textit{Analysis}: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

In this paper we present applicability of Page Ranking algorithm in web mining. This paper is organized as follows: Section II presents Web Mining Methodology, section III presents Page Rank algorithm, section IV presents Weighted Page rank Algorithm, section V presents HITS algorithm, section VI presents comparisons of these algorithms and finally in section we presents conclusion on this papers.

\section{WEB MINING METHODOLOGIES}

In this section we discuss web mining methodologies in short. Web mining, the application of machine learning (data mining) techniques to web based data for the purpose of learning or extracting
knowledge. Web mining methodologies can generally be classified into one of three distinct categories.

1) Web Content Mining
2) Web Structure Mining
3) Web Usage Mining

1) Web Content Mining: Web Content Mining is the process of retrieving the information from web document into more structure forms. It is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining.

2) Web Structure Mining: Web Structure Mining deals with the discovering and modelling the link structure of the web. This can help in discovering similarity between sites or discovering web communities.

3) Web Usage Mining: Web Usage Mining deals with understanding user behaviour in interacting with the web site. The aim is to obtain information that may assist web site recognition to better suit the user. The logs include information about the referring pages, user identification, time a user spends at a site and the sequence of pages visited.

II. PAGE RANKING ALGORITHM

With the increasing number of Web pages and users on the Web, the number of queries submitted to the search engines are also increasing rapidly. Therefore, the search engines need to be more efficient in its process. Web mining techniques are employed by the search engines to extract relevant documents from the web database and provide the necessary information to the users. The search engines become very successful and popular if they use efficient Ranking mechanism. Google search engine is very successful because of its PageRank algorithm.

Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document. If the search results are not displayed according to the user interest then the search engine will lose its popularity. So the ranking algorithms become very important.

Brin and Page developed PageRank algorithm during their Ph. D. at Stanford University based on the citation analysis [2]. PageRank algorithm is used by the famous search engine, Google. They applied the citation analysis in Web search by treating the incoming links as citations to the Web pages. However, by simply applying the citation analysis technique to the diverse set of Web documents did not result in efficient outcomes.

Therefore, PageRank provides a more advanced way to compute the importance or relevance of a Web page than simply counting the number of pages that are linking to it (called as “backlinks”). If a backlink comes from an “important” page, then that backlink is given a higher weighting than those backlinks coming from non-important pages. In a simple way, link from one page to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but the “importance” or the “relevance” of the ones that cast these votes as well.

Assume any arbitrary page A has pages T1 to Tn pointing to it (incoming link). PageRank can be calculated by the following (1).
\[ PR(A) = (1 - d) + d \left( \frac{PR(T1)}{C(T1)} + ... + \frac{PR(Tn)}{C(Tn)} \right) \] (1)

The parameter \( d \) is a damping factor, usually sets it to 0.85 (to stop the other pages having too much influence, this total vote is “damped down” by multiplying it by 0.85). \( C(A) \) is defined as the number of links going out of page \( A \).

The PageRanks form a probability distribution over the Web pages, so the sum of all Web pages’ PageRank will be one. PageRank can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web.

### III. WEIGHTED PAGE RANK ALGORITHM

Wenpu Xing and Ali Ghorbani [3] proposed a WeightedPageRank (WPR) algorithm which is an extension of the PageRank algorithm. This algorithm assigns a larger rank value to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as \( W_{in}(m,n) \) and \( W_{out}(m,n) \) respectively. \( W_{in}(m,n) \) as shown in (2) is the weight of link \((m,n)\) calculated based on the number of incoming links of page \( n \) and the number of incoming links of all reference pages of page \( m \).

\[ W_{in}^{in}(m,n) = \frac{l_n}{\sum_{p \in R(m)} l_p} \] (2)

\[ W_{out}^{out}(m,n) = \frac{o_n}{\sum_{p \in R(m)} o_p} \] (3)

Where \( In \) and \( Ip \) are the number of incoming links of page \( n \) and page \( p \) respectively. \( R(m) \) denotes the reference page list of page \( m \). \( W_{out}(m,n) \) is as shown in (3) is the weight of link \((m,n)\) calculated based on the number of outgoing links of page \( n \) and the number of outgoing links of all reference pages of \( m \). Where \( On \) and \( Op \) are the number of outgoing links of page \( n \) and \( p \) respectively. The formula as proposed by Wenpu et al. [3] for the WPR is as shown in (4) which is a modification of the PageRank formula.

\[ WPR(n) = (1-d) + d \sum_{m \in R(n)} WPR(m) W_{in}^{in}(m,n) W_{out}^{out}(m,n) \] (4)

### IV. HITS ALGORITHM

Kleinberg [4] identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject. Hubs and Authorities are shown in Fig. 1 [5]. In [4] Kleinberg says that a page may be a good hub and a good authority at the same time. This circular relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search). Hubs Authorities...
The HITS algorithm treats WWW as a directed graph $G(V,E)$, where $V$ is a set of Vertices representing pages and $E$ is a set of edges that correspond to links.

There are two major steps in the HITS algorithm. The first step, Sampling step, a set of relevant pages for the given query are collected i.e. a sub-graph $S$ of $G$ is retrieved which is high in authority pages. This algorithm starts with a root set $R$, a set of $S$ is obtained, keeping in mind that $S$ is relatively small, rich in relevant pages about the query and contains mostly of the good authorities.

The second step, Iterative step, find hubs and authorities using the output of the sampling step using (5) and (6).

$$H_p = \sum_{q \in I(p)} A_q$$  
$$A_p = \sum_{q \in B(p)} H_q$$

Where $H_p$ is the hub weight, $A_p$ is the Authority weight, $I(p)$ and $B(p)$ denotes the set of reference and referrer pages of page $p$. The page’s authority weight is proportional to the sum of the hub weights of pages that it links to, Kleinberg [6]. Similarly, a page’s hub weight is proportional to the sum of the authority weights of pages that it links to.

HITS algorithm has some constraints which are explain below [7]:

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- Automatically generated links: HITS gives equal importance for automatically generated links which may not produce relevant topics for the user query.
Efficiency: HITS algorithm is not efficient in realtime. HITS was used in a prototype search engine called Clever[7] for an IBM research project. Because of the above constraints HITS could not be implemented in a real time search engine.

V. COMPARISON OF ALGORITHMS

Table I shows the comparison [8] of all the algorithms discussed above. The main criteria used for comparison are mining techniques used, working method, input parameters, complexity, limitations and the search engine using the algorithm. Among all the algorithms, PageRank and HITS are most important ones. PageRank is the only algorithm implemented in the Google search engine. HITS is used in the IBM prototype search engine called Clever. A similar algorithm is used in the Teoma search engine and later it is acquired by Ask.com. HITS cannot be implemented directly in a search engine due to its topic drift and efficiency problem. That is the reason we have taken PageRank algorithm and implemented in a Java program.

Table I. Comparison of algorithms [8]

<table>
<thead>
<tr>
<th>Algorithms/Criteria</th>
<th>PageRank</th>
<th>Weighted Page Rank</th>
<th>HITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining techniques used</td>
<td>WSM</td>
<td>WSM</td>
<td>WSM and WCM</td>
</tr>
<tr>
<td>Working</td>
<td>Computes scores at indexing time. Results are sorted according to importance of pages. Backlinks</td>
<td>Computes scores at indexing time. Results are sorted according to page importance. Backlinks, Forward links</td>
<td>Computers hub and authority scores of n highly relevant pages on the fly. Backlinks, Forward links and content &lt; O(log N)</td>
</tr>
<tr>
<td>I/P Parameters</td>
<td>Google</td>
<td>&lt; O(log N)</td>
<td>Topic drift and efficiency problem Clever</td>
</tr>
<tr>
<td>Complexity</td>
<td>Query independent</td>
<td>Query independent</td>
<td></td>
</tr>
<tr>
<td>Limitations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search Engine</td>
<td>Research model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper we studied basic of web mining and its methodology. Special purpose of this paper is that we explain three most important algorithms of web mining Page Rank, Weighted Page Rank and HITS algorithm with formulas.

We also explains comparisons of these algorithms with different criteria such as mining techniques, I/P parameters, complexity, limitations and search engine used.

REFERENCES