ISSN: 2229 – 6948 (ONLINE)
DOI: 10.21917/ijct.2011.0039

WEB STRUCTURE MINING USING PAGERANK, IMPROVED PAGERANK – AN OVERVIEW

V. Lakshmi Praba[1] and T. Vasantha[2]
[1]*Department of Computer Science, Government Arts College for Women, Tamil Nadu, India*
E-mail: vlakshmipraba@rediffmail.com
[2]*Manonmaniam Sundaranar University, Tamil Nadu, India*
E-mail: vasanthasankarganesh@gmail.com

Abstract
Web Mining is the extraction of interesting and potentially useful patterns and information from Web. It includes Web documents, hyperlinks between documents, and usage logs of web sites. The significant task for web mining can be listed out as Information Retrieval, Information Selection / Extraction, Generalization and Analysis. Web information retrieval tools consider only the text on pages and ignore information in the links. The goal of Web structure mining is to explore structural summary about web. Web structure mining focusing on link information is an important aspect of web data. This paper presents an overview of the PageRank, Improved Page Rank and its working functionality in web structure mining.

Keywords:
Web Structure Mining, Improved PageRank

1. INTRODUCTION

Web mining is the application of data mining and knowledge discovery techniques to extract information from Web. Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) are the three knowledge discovery domains in Web mining [3]. Web content mining is the process of extracting knowledge from the content of documents - including text, image, audio, video, etc. Research in web content mining includes resource discovery from the web, clustering such as document based clustering and Link based clustering, and information extraction from web pages. Web structure mining is the process of inferring knowledge from the Web and links between references and referents in the Web. Web usage mining involves the automatic discovery of user access patterns from one or more Web servers [3].

2. WEB STRUCTURE MINING

Web structure mining deals with the structure of the hyperlinks within the web itself. In this paper the structure of web as a directed graph with web pages and hyperlinks, where the web pages correspond to nodes and hyperlinks correspond to edges connecting the related pages has been focused. An overview of the two kinds of web structure namely**,** *Link Structure, Document Structure* has been considered.

2.1 LINK STRUCTURE

Link Structure is based on hyperlink structure. Hyperlink Structure is a structural unit that connects a place in a Web page to a different place, either within the same Web page or on a different Webpage. A hyperlink that connects to a different place within the same web page is called as *Intra-Page Hyperlink*, and hyperlink that connects a particular place in another web page is called as *Inter-Page Hyperlink* [3].

2.2 DOCUMENT STRUCTURE

Document Structure represents the designer's view of the content organization within a web page. The content represented in the arrangement of HTML and XML tags within a page is called as *Intra- page structure*. Both HTML and XML documents can be represented in tree structure format.

3. WEB STRUCTURE ALGORITHMS

Web structure mining discovers the structure of web document as well as link structure of the hyperlinks. The hyperlink structure of the web is valuable for locating information. Based on the topology of the hyperlinks, it will categorize the web pages and generate the information, such as the similarity and relationship between different web sites. WSM has two influential hyperlink based search algorithms – PageRank Algorithm and Improved PageRank Algorithm.

3.1 PAGERANK ALGORITHM

PageRank was presented by Sergey Brin and Larry Page. The search engine Google was built based on this algorithm. PageRank has emerged as the dominant link analysis model for Web search, partly due to its query-independent evaluation of Web pages and its ability to combat spamming [4].

PageRank suggest that a link from page *u* to *v* implies the author of *u* recommends taking a look at page *v*. A page has a high PageRank value if there are many pages that point to it. PageRank is a static ranking of Web pages in the sense that a PageRank value is computed for each page off-line and it does not depend on search queries. The PageRank Algorithm represents the structure of the Web as a matrix, and PageRank values as a vector. The PageRank vector is derived by computing matrix-vector multiplications repeatedly. The sum of all PageRank values should be one during the computation [4].

3.2 PAGERANK COMPUTATION

The PageRank Algorithm is the most popular method to compute the PageRank value of web sites. The PageRank computation based on the concepts of back links. Considering *d* as a web page, the set of web pages that points to *d* are called as back links of *d* , denoted by *Bd*. Set of pages pointed by *d* are called as forward links, denoted by *Fd*. The formula to calculate the pagerank is as follows:

$$PR(m)= \sum_{n \varepsilon Bm} PR(n)/N(n) \quad (1)$$

The web can be viewed as a directed graph whose nodes are the pages or documents and the edges are the hyperlinks between them. This directed graph structure in the web is called as web graph. A graph can be represented as G*(V,E)* consist of set of vertices and set of edges are denoted by *V(G)* and *E(G)* respectively. Let us assume the Hyperlink Structure for a web graph of eight pages (Fig.1), and the corresponding matrix for the designed Hyperlink Structure is represented in Fig.2. In this Hyperlink Structure pages are denoted by rectangular shape. The links are denoted by directed line from one page to another. Every page in this Hyperlink Structure has at least one outgoing edge.

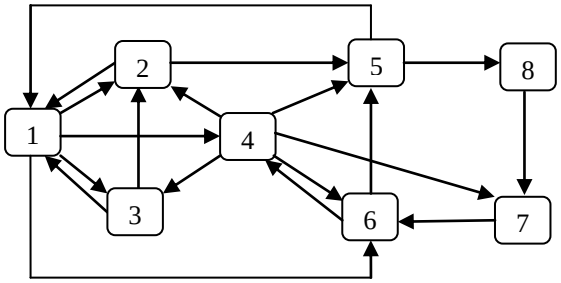

Fig.1. Hyperlink Structure for a web graph of eight pages

M

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	**0**	**1 / 2**	**1 / 2**	**0**	**1 / 2**	**0**	**0**	**0**
2	**1 / 4**	**0**	**1 / 2**	**1 / 5**	**0**	**0**	**0**	**0**
3	**1 / 4**	**0**	**0**	**1 / 5**	**0**	**0**	**0**	**0**
4	**1 / 4**	**0**	**0**	**0**	**0**	**1 / 2**	**0**	**0**
5	**0**	**1 / 2**	**0**	**1 / 5**	**0**	**1 / 2**	**0**	**0**
6	**1 / 4**	**0**	**0**	**1 / 5**	**0**	**0**	**1**	**0**
7	**0**	**0**	**0**	**1 / 5**	**0**	**0**	**0**	**1**
8	**0**	**0**	**0**	**0**	**1 / 2**	**0**	**0**	**0**

Fig.2. Matrix for Hyperlink Structure

Entries in the square matrix are determined by link between the pages. If there is no link between the pages then the value of the entry is zero otherwise its value is 1/P where P is number of outgoing edges[1],[4].

Each row of the square matrix indicate back link of every node and each column of the square matrix indicate forward link of every node. Using the values of matrix the rank can be calculated for all pages in the web graph. Back link of page 1 is represented by Ist row of the matrix and forward link of page 1 is represented by Ist column of the matrix. PageRank value of each page is obtained from summing the rank of Non zero values of its corresponding row in matrix *M*.

An initial rank matrix $Rank_0$ is defined as a N x 1 where N is the total number pages in the web graph All entries in $Rank_0$ are set to 1/N. $Rank_0$ = (1/N, 1/N, 1/N, 1/N, 1/N, 1/N, 1/N, 1/N)^T ,T stands for transposition. Rank can be obtained from the product of square matrix M and Rank. such as $Rank_i = M \times Rank_{i-1}$. To determine the PageRank value, a number of iterations are carried out until it converges to a fixed point. The converged $Rank_i$ is the PageRank values of all pages[5],[6].

Compute new matrix M' by applying dampening factor *d* with square matrix *M*. where *d* is the probability that user will link to a random page and (*1-d*) is the probability that user will follow the links consecutively, to trace the web. M' is expressed as Eq. (2).

$$M' = (1 - d)M + d\left[\frac{1}{N}\right] NXN \quad (2)$$

where *N* is the total number of pages and [1/*N*] *N*×*N* be an *N*×*N* square matrix in which all entries are *(1/N)*. The constant *d* is less than one. According to the literature review, many researches (Sung Jin Kim and Sang Ho Lee, Hung-Yu Kao and Seng-Feng Lin) have suggested normally *d* is set to 0.15. Hyperlink Structure for Fig.1 illustrates no Dangling Pages and no RankSink problem. M′ constructed for the Fig.1 using Eq. (2) is given below.

M′

	1	2	3	4	5	6	7	8
1	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
2	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
3	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
4	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
5	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
6	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1)+d(1/N)	(1-d)(0)+d(1/N)
7	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1)+d(1/N)
8	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)

When N = 8 and d = 0.15, based on the value of N and d, Matrix M′ calculated with initial rank value is illustrated in Fig.3.

M′ = x

0.0188	0.4437	0.4437	0.0188	0.4437	0.0188	0.0188	0.0188
0.2313	0.0188	0.4437	0.1887	0.0188	0.0188	0.0188	0.0188
0.2313	0.0188	0.0188	0.1887	0.0188	0.0188	0.0188	0.0188
0.2313	0.0188	0.0188	0.0188	0.0188	0.4437	0.0188	0.0188
0.0188	0.4437	0.0188	0.1887	0.0188	0.4437	0.0188	0.0188
0.2313	0.0188	0.0188	0.1887	0.0188	0.0188	0.8687	0.0188
0.0188	0.0188	0.0188	0.1887	0.0188	0.0188	0.0188	0.8687
0.0188	0.0188	0.0188	0.0188	0.4437	0.0188	0.0188	0.0188

Rank = $\begin{pmatrix} 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \end{pmatrix}$

Fig.3. Calculated Matrix Values of M′ and initial rank value

The Eq. (3) is used to compute Page Rank values of all pages as given below:

$$Rank_{i+1} = M' X \ Rank_i \quad (3)$$

The PageRank calculation equations for all the eight pages are given below.

Page1- PageRank Calculation:

$$\begin{aligned} PR[1] &= ((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &\quad +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) + ((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &= ((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) + ((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) \\ &= 0.0023 +0.0555 +0.0555 +0.0023 +0.0555+0.0023 +0.0023 +0.0023 = 0.1781 \end{aligned} \quad (4)$$

Page2- PageRank Calculation:

$$\begin{aligned} PR[2] &= ((1-d)(\tfrac{1}{4})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{5})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &\quad +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &= ((1-0.15)(\tfrac{1}{4})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{5})+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0) + 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) + ((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) \\ &= 0.0289+0.0023 +0.0555+0.0236+0.0023 +0.0023 +0.0023 +0.0023 = 0.1197 \end{aligned} \quad (5)$$

Page3- PageRank Calculation:

$$\begin{aligned} PR[3] &= ((1-d)(\tfrac{1}{4})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{5})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &\quad +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &= ((1-0.15)(\tfrac{1}{4})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{5})+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) + ((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) \\ &= 0.0289+0.0023 +0.0023+0.0236+0.0023 +0.0023 +0.0023 +0.0023 = 0.0666 \end{aligned} \quad (6)$$

Page4 - PageRank Calculation:

$$\begin{aligned} PR[4] &= ((1-d)(\tfrac{1}{4})+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) + ((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &\quad +((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &= ((1-0.15)(\tfrac{1}{4})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) \\ &= 0.0289+0.0023 +0.0023+0.0023 +0.0023 +0.0555+0.0023 +0.0023 = 0.0984 \end{aligned} \quad (7)$$

Page5 - PageRank Calculation:

$$\begin{aligned} PR[5] &= ((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{5})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{2}) \\ &\quad +d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) \\ &= ((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{5})+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{2})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8})) (\tfrac{1}{8}) \\ &= 0.0023 +0.0555+0.0023+0.0236+0.0023 +0.0555+0.0023 +0.0023 = 0.1462 \end{aligned} \quad (8)$$

Page6 - PageRank Calculation:

$$\begin{aligned} PR[6] &= ((1-d)(\tfrac{1}{4})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(\tfrac{1}{5})+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &\quad +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(1)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \\ &= ((1-0.15)(\tfrac{1}{4})+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{5})+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(1)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &= 0.0289+0.0023 +0.0023+0.0236+0.0023 +0.0023 +0.1086+0.0023 = 0.1728 \end{aligned} \quad (9)$$

Page7 - PageRank Calculation:

$$\begin{aligned} PR[7] &= ((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) +((1-d)(\tfrac{1}{5})+d(\tfrac{1}{N})) (\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0) \\ &\quad +d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N})) (\tfrac{1}{N}) +((1-d)(1)+d(\tfrac{1}{N})) (\tfrac{1}{N}) \\ &= ((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(\tfrac{1}{5})+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &\quad +((1-0.15)(0) + 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(1)+ 0.15(\tfrac{1}{8}))(\tfrac{1}{8}) \\ &= 0.0023 +0.0023 +0.0023+0.0236+0.0023 +0.0023 +0.0023 +0.1086= 0.1463 \end{aligned} \quad (10)$$

Page8 - PageRank Calculation:

$$PR[8] = ((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) + ((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) +((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N})+((1-d)(0)+d(\tfrac{1}{N}))(\tfrac{1}{N}) +((1-d)(1)+d(\tfrac{1}{N}))(\tfrac{1}{N}) \quad (11)$$

$$= ((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8}) +((1-0.15)(\tfrac{1}{2})+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})+((1-0.15)(0)+0.15(\tfrac{1}{8}))(\tfrac{1}{8})$$

$$= 0.0023 +0.0023 +0.0023+0.0023 +0.0555+0.0023 +0.0023 +0.0023 = 0.0719$$

By repeatedly doing the PageRank calculations using the Eqs. (4-11), the PageRank values are computed with no Dangling page and no RankSink[1][12] are given in Table.1.

Table.1. PageRank with no Dangling page and no RankSink

Iteration	Page1	Page2	Page3	Page4	Page5	Page6	Page7	Page8	TotPR	PRLoss
1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	1.0000	0.0000
2	0.1781	0.1197	0.0666	0.0984	0.1462	0.1728	0.1463	0.0719	1.0000	0.0000
3	0.1601	0.1016	0.0733	0.1300	0.1598	0.1976	0.0966	0.0809	1.0000	0.0000
4	0.1610	0.1060	0.0749	0.1368	0.168	0.1570	0.1096	0.0867	1.0000	0.0000
5	0.1671	0.1080	0.0762	0.1197	0.1538	0.1694	0.1157	0.0902	1.0000	0.0000
6	0.1624	0.1070	0.0746	0.1262	0.1570	0.1729	0.1157	0.0841	1.0000	0.0000
7	0.1626	0.1064	0.0747	0.1267	0.1592	0.1731	0.1117	0.0855	1.0000	0.0000
8	0.1634	0.1066	0.0749	0.1269	0.1591	0.1698	0.1130	0.0864	1.0000	0.0000
9	0.1635	0.1069	0.075	0.1256	0.1578	0.171	0.1138	0.0864	1.0000	0.0000
10	0.1631	0.1067	0.0749	0.1262	0.1582	0.1715	0.1135	0.0858	1.0000	0.0000
11	0.1632	0.1067	0.0749	0.1263	0.1585	0.1714	0.1131	0.0860	1.0000	0.0000
12	0.1633	0.1067	0.0749	0.1263	0.1584	0.1711	0.1133	0.0861	1.0000	0.0000
13	0.1633	0.1067	0.0749	0.1261	0.1583	0.1712	0.1134	0.0861	1.0000	0.0000
14	0.1632	0.1067	0.0749	0.1262	0.1583	0.1713	0.1134	0.086	1.0000	0.0000
15	0.1632	0.1067	0.0749	0.1262	0.1584	0.1712	0.1133	0.086	1.0000	0.0000
16	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.0861	1.0000	0.0000
17	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
18	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
19	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
20	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
21	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
22	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
23	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
24	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000
25	0.1632	0.1067	0.0749	0.1262	0.1583	0.1712	0.1133	0.086	1.0000	0.0000

3.2.1 RankSink:

RankSink is one of the problems of Page Rank Algorithm. Consider number of web pages that point internally to each other but do not point to other web pages (that is a loop). Assume there are some web pages that point to one of pages in the loop. Then, during the iteration, the loop accumulates PageRank values but never distributes any PageRank values .The loop forms a sort of trap, which is called as RankSink. Fig.4 demonstrates a simple example of loop which consists of three web pages and it act as RankSink [5]. Due to RankSink problem, pages in the loop are likely to have higher PageRank values than the existence [10], [11].

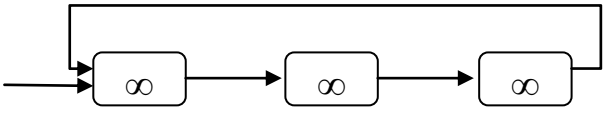

Fig.4. Loop which act as Rank Sink

For the illustration shown in Fig.1, suppose the link from pages 1,2,4,5 to pages 4,5,3,1 respectively and a link from 1 to 6 are removed, the remaining links are represented in Fig.5. It illustrates a RankSink structure for a web graph with eight pages. Pages 1, 2, and 3 form a group 1, and the rest of the pages form a group 2. Note that a link from page 4 to page 2, which is an only link from group 2 to group 1. Existing PageRank Algorithm overcome the RankSink problem [1].

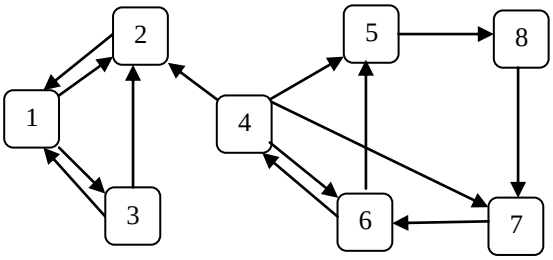

Fig.5.RankSink structure for eight page web

3.2.2 Dangling Page:

Major problem of original PageRank Algorithm is Dangling page. If a page in web graph has no forward link then the page is called as Dangling page. In Fig.6 the page *5* is a Dangling page as it has no outgoing edge, and the 5^{th} column of corresponding matrix *M* is called as dangling column. It is represented as a zero vector (Fig.7) . In general a matrix *M* is irreducible if and only if a directed graph is strongly connected. All columns in an irreducible matrix *M* have norms of value one. Here norm of 5^{th} column of matrix *M is* zero. As a consequence of Dangling page, $Rank_i$ loses a part of its norm continuously during the iteration process. This event is called as *norm-leak*[1],[10].

3.3 IMPROVED PAGERANK ALGORITHM

Fig.6. demonstrates the Hyperlink Structure with Dangling page and its matrix is given Fig.7.

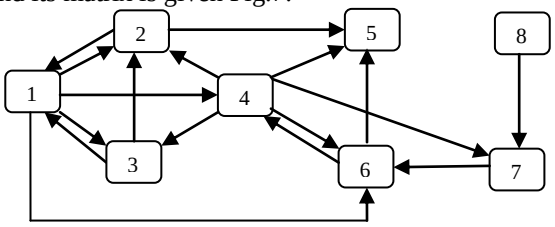

Fig.6. Hyperlink Structure with Dangling page

M

	1	2	3	4	5	6	7	8
1	**0**	**1 / 2**	**1 / 2**	**0**	**0**	**0**	**0**	**0**
2	**1 / 4**	**0**	**1 / 2**	**1 / 5**	**0**	**0**	**0**	**0**
3	**1 / 4**	**0**	**0**	**1 / 5**	**0**	**0**	**0**	**0**
4	**1 / 4**	**0**	**0**	**0**	**0**	**1 / 2**	**0**	**0**
5	**0**	**1 / 2**	**0**	**1 / 5**	**0**	**1 / 2**	**0**	**0**
6	**1 / 4**	**0**	**0**	**1 / 5**	**0**	**0**	**1**	**0**
7	**0**	**0**	**0**	**1 / 5**	**0**	**0**	**0**	**1**
8	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Fig.7. Matrix for Hyperlink Structure with Dangling page

Existing PageRank Algorithm can't solve the norm-leak problem, because the norms of $Rank_i$ and $Rank_{i+1}$ never converged to a fixed point.

3.4 A NEW MATRIX, M^+

The occurrence of *norm_leak* can be overcome by the Improved PageRank using new matrix M^+. The dangling column of *M* is replaced by[1/N]*NX1* in order to compute matrix M^+. It can be expressed as follows:

$$M^+ = M + \sum_{P \in D} < \frac{1}{N} > NXN, P \qquad (12)$$

where *D* is a set of Dangling pages. Let *<1/N>NXN,p* be an *NXN* matrix in which entry m_{ij} is *(0)* if *j* is not equal to *p* otherwise m_{ij} is (1/N). The *p*th column of *<1/N>NXN,p* is exactly [1/*N*]*NX1*[1],[7].

At the end of processing the Dangling page has complete set of outgoing edges to all nodes including it. The matrix M^+ implies that each Dangling page has a complete set of edges which are as follows:

M^+

	1	2	3	4	5	6	7	8
1	**0**	**1 / 2**	**1 / 2**	**0**	**1/N**	**0**	**0**	**0**
2	**1 / 4**	**0**	**1 / 2**	**1 / 5**	**1/N**	**0**	**0**	**0**
3	**1 / 4**	**0**	**0**	**1 / 5**	**1/N**	**0**	**0**	**0**
4	**1 / 4**	**0**	**0**	**0**	**1/N**	**1 / 2**	**0**	**0**
5	**0**	**1 / 2**	**0**	**1 / 5**	**1/N**	**1 / 2**	**0**	**0**
6	**1 / 4**	**0**	**0**	**1 / 5**	**1/N**	**0**	**1**	**0**
7	**0**	**0**	**0**	**1 / 5**	**1/N**	**0**	**0**	**1**
8	**0**	**0**	**0**	**0**	**1/N**	**0**	**0**	**0**

Fig.8. Calculated Matrix Values of M^+

Improved PageRank Algorithm is used to overcome the Dangling Page problem by applying dampening factor *d* in the square matrix M^+ in order to obtain another matrix *M**.The *M** can be proposed as

$$M^* = (1 - d)\, M^+ + d[\tfrac{1}{N}]_{NXN} \qquad (13)$$

By using the Eq. (13), M^* is constructed for Fig.6 is follows

M*

	1	2	3	4	5	6	7	8
1	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
2	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
3	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
4	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
5	(1-d)(0)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(1/2)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)
6	(1-d)(1/4)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1)+d(1/N)	(1-d)(0)+d(1/N)
7	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/5)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1)+d(1/N)
8	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(1/N)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)	(1-d)(0)+d(1/N)

Using Improved PageRank Algorithm, calculated PageRank values for each page is illustrate in Table.2 .It shows that the result might converge to a fixed point, In an Improved PageRank Algorithm, the sum of all PageRank values are always maintained to be one [8],[9].

Table.2. Improved Page Rank Overcome the Dangling Page

Iteration	Page1	Page2	Page3	Page4	Page5	Page6	Page7	Page8	TotPR	PRLoss
1	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	1.0000	0.0000
2	0.1383	0.1330	0.0798	0.1117	0.1595	0.1861	0.1595	0.0320	1.0000	0.0000
3	0.1261	0.1180	0.0841	0.1442	0.1903	0.2197	0.0819	0.0357	1.0000	0.0000

4	0.1249	0.1260	0.0903	0.1591	0.2070	0.1599	0.0938	0.0390	1.0000	0.0000
5	0.1327	0.1327	0.0943	0.1352	0.1893	0.1741	0.1009	0.0407	1.0000	0.0000
6	0.1354	0.1301	0.0900	0.1410	0.1922	0.1758	0.0965	0.0389	1.0000	0.0000
7	0.1328	0.1302	0.0919	0.1427	0.1932	0.1739	0.0962	0.0392	1.0000	0.0000
8	0.1337	0.1308	0.0917	0.1414	0.1928	0.1735	0.0968	0.0393	1.0000	0.0000
9	0.1338	0.1307	0.0917	0.1414	0.1926	0.1740	0.0967	0.0392	1.0000	0.0000
10	0.1337	0.1306	0.0917	0.1416	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
11	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
12	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
13	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
14	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
15	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
16	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
17	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
18	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
19	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
20	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
21	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
22	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
23	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
24	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000
25	0.1337	0.1307	0.0917	0.1415	0.1927	0.1738	0.0966	0.0392	1.0000	0.0000

4. EVALUATION

PageRank Algorithm and the Improved PageRank Agorithm are implemented in C language and executed in Intel(R) Pentium (R) Dual CPU (200 Ghz) machine with 2GB of RAM. The number of nodes and the number of links are entered as input for both algorithms and the PageRank values are obtained as output. The elapsed times for PageRank Algorithm and Improved PageRank Algorithm to compute PageRank values are 0.44 ms and 0.49 ms respectively. For the demonstrated web graph, Fig.6 which has one Dangling Page and no RankSink, the obtained results are illustrates in Fig.9. The algorithm is executed for 25 iterations and it is observed that the $Rank_i$ and $Rank_{i+1}$ never converged for a web graph with Dangling pages.

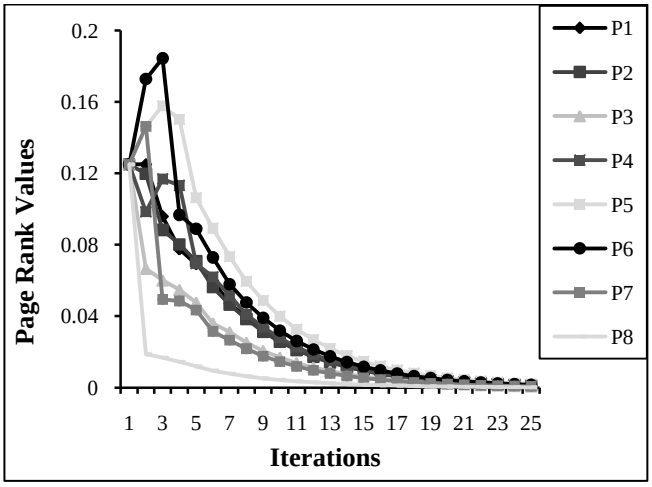

Fig.9. Result of PageRank Algorithm

Fig.10 shows the result of the Improved PageRank Algorithm. The Improved PageRank Algorithm is executed for same web graph. It is observed that both $Rank_i$ and $Rank_{i+1}$ converged at the earlier stage of the iteration itself, which then remain unchanged.

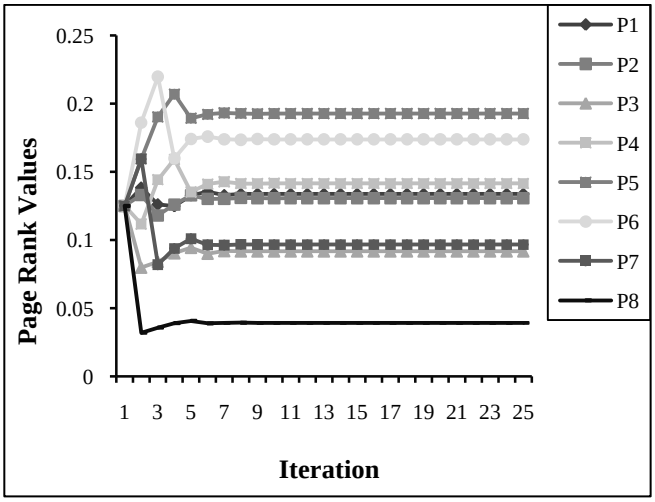

Fig.10. Result of Improved PageRank Algorithm

5. CONCLUSION

The main purpose of this study is to explore the hyperlink structure and to understand the web graph. This study focused on Improved PageRank Algorithm which solves the *norm-leak* phenomenon. In the Improved PageRank Algorithm, pages have the exact PageRank values. This study included a number of pre-processing operations too. The evaluation results clearly indicated that the rank of Improved PageRank Algorithm always converged to a fixed point which is not the case with PageRank Algorithm. This Improved PageRank Algorithm finds applications in search engines.

Further research can be focused by analyzing the links between web sites. Algorithms like HITS (Hyperlink Induced

Topic Search) and Weighted PageRank can also be evaluated in similar line.

REFERENCES

[1] F.Crestani,M.Girolami and C..J.Van Rijsbergen (Eds.): ECIR 2002, LNCS 2291, pp.73-85, 2002 ©Springer-Verlag Berlin Heidelberg.

[2] Chakrabarti, S., B. Dom, D. Gibson, J. Kleinberg and R. Kumar *et al*., 1999, "Mining the link structure of the world wide web", IEEE Comput., Vol. 32: pp. 60-67.

[3] T. Srivastava ,Prasanna Desikan, Vipin Kumar, 2005, "Web Mining – Concept, Applications and Research Directions" , Studies in Fuzziness and Soft Computing, Vol. 180, pp. 275 – 307.

[4] Bing Liu, 2007, "Web Data Mining, Exploring Hyperlinks,Contents, and Usage Data", ©Springer-Verlag Berlin Heidelberg.

[5] K. Bharat and M. Henzinger, 1998, "Improved Algorithms for Topic Distillation in Hyperlinked Environments", Proceedings of the 21st ACM SIGIR Conference, pp.104-111.

[6] S.Brin and L.Page, 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceedings of WWW7, pp.107-117.

[7] J.Dean and M. R. Henzinger, 1999, "Finding Related Web Pages in the World Wide Web", Proceedings of WWW8, pp.1467-1479.

[8] D.Gibson, J.Kleinberg, and P. Raghavan, 1998, "Inferring Web Communities from Link Topology", Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, pp.225-234.

[9] Y.Ng, A.X.Zheng, and M.I. Jordan, 2001, "Stable Algorithms for Link Analysis", Proceedings of the 24th ACM SIGIR Conference, pp.258-266.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, 1998, "The PageRank Citation Ranking: Bringing Order to the Web", unpublished manuscript, Stanford University.

[11] T.H.Haveliwala, 1999, "Efficient Computation of PageRank, unpublished manuscript", Stanford University.

[12] http://www.google-pagerank-improve.com.