Google Page Rank Algorithm and beyond

Submitted by
Digbalay Bose
Roll no:143070026
Mtech 1st Year Control and Computing
Guided by: Prof. Debasattam Pal

Department of Electrical Engineering
Indian Institute of Technology, Bombay
November, 2014

Abstract

In this article, a brief insight has been given regarding the famous page rank algorithm , which is used by the Google search engine. The algorithm operates on the principle of assigning fixed scores i.e. Page Rank values to web pages which are connected by hyperlinks. These scores help in developing the idea of importance of web pages , which results in higher Page Rank values for web pages visited more frequently.A brief insight into the page rank algorithm has been given followed by analysis using Perron Frobenius theorem of irreducible matrices and Banach fixed point theorem for contraction mapping. Further, the idea of aggregation based page ranking system is discussed ,followed by another application of page ranking i.e. ranking of journals.

1 Introduction

The Page Rank algorithm [1] which is the bedrock of the Google search engine [3] was developed in the late 1990s. It operates on the principle of assigning importance scores to the web pages based on the structure of the hyper links in the web. Those web pages , which have greater number of incoming links are more important than other web pages. The analysis in this article is motivated by a graph model of the web i.e. hyperlink model , where pages represent nodes and hyperlinks indicate edges.The graph modelling of the web enables the presentation of web surfing activity as a random walk . The simple random walk model into was further developed into a teleportation model to incorporate the idea of random jumps among webpages.
The article is constructed as follows:Section 2 provides a brief idea about the hyperlink model along with a Markov chain based analysis of the model .This is followed by the improved teleportation model in section 3. Section 4 includes the explanation of Perron Frobenius theorem and its application in PageRank analysis. Further, the idea of contraction mapping, utilized in Banach Fixed Point theorem [6] , is investigated in the analysis of Page Ranking system in section 5. A paradigm of page ranking i.e. aggregation based page ranking [2] is also discussed in section 6 followed by application of page ranking in the ordering of journals.

2 Hyperlink model

In this section a graph based model for the web has been considered for analysis, where the nodes represent the pages and the edges indicate presence of hyperlinks. The graph considered is directed in nature and denoted by $G = (V, E)$, where V is the set of vertices and E is the set of edges. The directed nature of the graph helps in distinguishing between incoming and outgoing links i.e. an edge (i, j) indicates that a link exists from vertex i to vertex j with the direction from i to j. Thus this link is outgoing in nature for vertex i and incoming for vertex j. The number of incoming links associated with a vertex i is helpful in determining the importance of the vertex i or page i.
The representation of the graph in terms of matrix notation is given by considering a set $Linked_i$, where $Linked_i$ consists of all pages which have outgoing links to page i and is denoted by $Linked_i = \{j : (j, i) \in E\}$. The entries of the resulting matrix usually called hyperlink matrix H are given by :-

$$H_{ij} = \begin{cases} \frac{1}{O_j} & j \in Linked_i \\ 0 & otherwise \end{cases} \tag{1}$$

Here O_j determines the number of outgoing links from node j. The entries of the matrix H satisfy the following relations:

$$H_{ij} \geq 0 \tag{2}$$

$$\sum_{i=1}^{N} H_{ij} = 1 \tag{3}$$

Here N is the number of nodes or webpages in the model considered. An example is considered for illustrating the hyperlink model in which there are 6 nodes labelled A, B, C, D, E, F with links given by directed edges. The graph for the hyperlink model is given in figure(1).
Considering the example of the following hyperlink model, the resulting hyperlink matrix, which is a

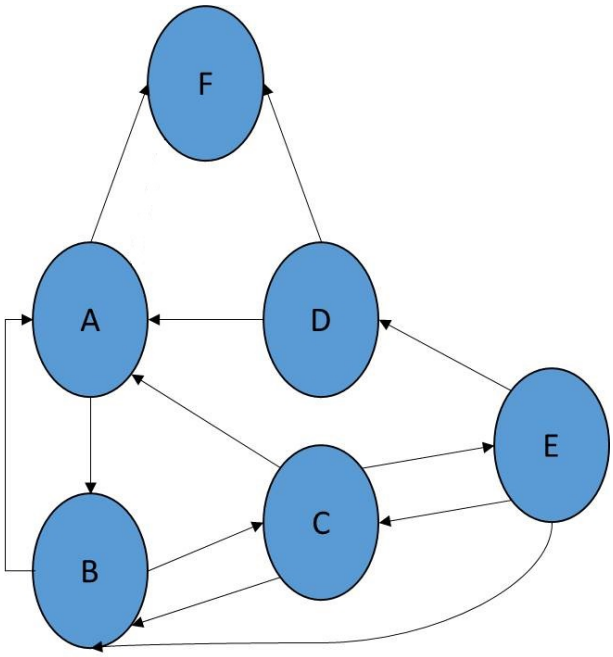

Figure 1: **Graph modelling of web having 6 nodes i.e. 6 webpages(Here node F has no outgoing edge)**.

square matrix of size 6×6 , is given as:-

$$H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix} \tag{4}$$

The presence of all zeroes in the last column indicate the absence of outgoing links from node F. Examples include pages containing pdf files and images.To alleviate this problem an outgoing link is added from node F to node A, resulting in change in the entries of the 6 th column . The final updated hyperlink matrix is given by:-

$$H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix} \tag{5}$$

The resulting graph for the hyperlink model obtained by introducing an edge from node F is given in figure (2).

2.1 Analysis of hyperlink model based on Markov chain properties

A Markov process is characterized by the property that the transition from one state to another in a given time interval depends on the two states only. If the number of states is finite,then the Markov process is known as Markov chain. In this subsection the idea of web surfing as a markov chain [4] will be considered, with special emphasis on the convergence of the Page Rank values of diiferent webpages. For this purpose , a random variable X_n is considered, where X_n for the above mentioned hyperlink model can take any values from the set of webpages i.e. A, B, C, D, E, F.

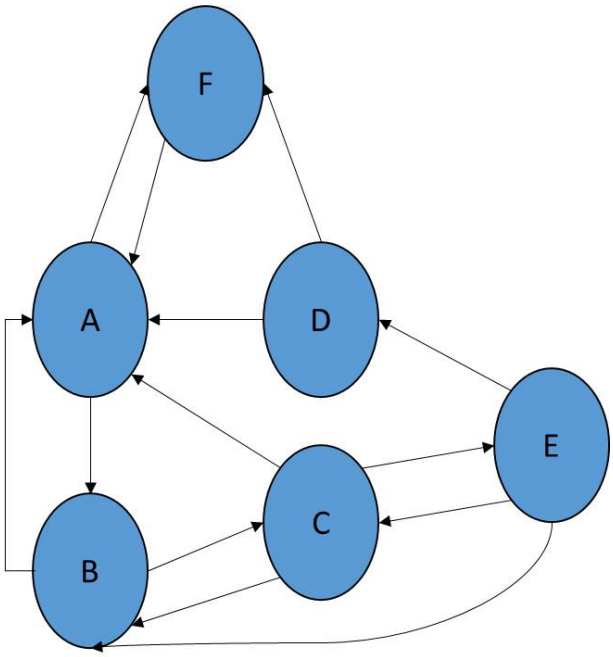

Figure 2: **Graph modelling of web having 6 nodes i.e. 6 webpages(Here node A has an incoming edge from node F)**.

X_n helps in determining the current location after n steps with each entry of the hyperlink matrix H i.e. h_{ij} denoting the the conditional probability $h_{ij} = P(X_n = i|X_{n-1} = j)$. If the page rank values at the kth instant are given by $x(k)$,then their updating is performed using the following rule :

$$x(k+1) = Hx(k). \tag{6}$$

Since the hyperlink matrix entries are probability values , after n traversals the coressponding conditional probabilities are listed in the matrix H^n. For an initial set of probabilities(i.e. page rank values for the six webpages A, B, C, D, E, F) given by $P^0 = [1/3, 1/3, 0, 0, 1/3, 0]^t$,the probability vector after the first step is given by

$$p^1 = H \cdot P^0 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1667 \\ 0.2778 \\ 0.2778 \\ 0.1111 \\ 0 \\ 0.1667 \end{bmatrix}$$

For n steps the probability vector P can be calculated as $p^n = H \cdot p^{n-1}$. The hyperlink matrix H has some interesting properties as listed below.

Theorem 2.1. *The hyperlink matrix H has 1 as an eigenvalue*

Proof. For the hyperlink matrix H considered, its transpose H^t also has the same set of eigen values. Defining u to be a vector $\in R^N$ (Hyperlink Matrix is of size $N \times N$) , such that all its entries are equal to 1, we have

$$u = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Then $H^t u$ is given as :

$$\begin{bmatrix} H_{11} & H_{21} & H_{31} & \cdots & H_{N1} \\ H_{12} & H_{22} & H_{32} & \cdots & H_{N2} \\ H_{13} & H_{23} & H_{33} & \cdots & H_{N3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ H_{1N} & H_{2N} & H_{3N} & \cdots & H_{NN} \end{bmatrix} . \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} H_{11} + H_{21} + H_{31} + \cdots + H_{N1} \\ H_{12} + H_{22} + H_{32} + \cdots + H_{N2} \\ \vdots \\ \vdots \\ H_{1N} + H_{2N} + H_{3N} + \cdots + H_{NN} \end{bmatrix}$$

Since the hyperlink matrix H has the property $\sum_{i=1}^{N} H_{ij} = 1$, then each entry of $H^t u$ is 1. So, we have $H^t u = u$, which implies that H^t has 1 as an eigenvalue and u as the eigen vector. Thus, H has an eigenvalue 1. □

Theorem 2.2. *All the eigenvalues of H satisfy the relation* $|\lambda| \leq 1$ *i.e. 1 is the largest eigenvalue.*

Proof. Considering $\lambda \in C$ to be an eigenvalue of hyperlink matrix H, then λ is also an eigenvalue of H^T .If v is the eigen vector of H^T coressponding to eigen value λ , then the following relation holds

$$H^T v = \lambda v \tag{7}$$

Considering the index m such that $|v_j| \leq |v_m|$, $\forall j$, $1 \leq j \leq N$, then we have from (7) for the mth component on both sides

$$\sum_{j=1}^{N} H_{jm} v_j = \lambda v_m \tag{8}$$

Thus the following can be obtained

$$|\lambda v_m| = |\lambda| \cdot |v_m| = |\sum_{j=1}^{N} H_{jm} v_j| \leq \sum_{j=1}^{N} H_{jm} |v_j| \leq \sum_{j=1}^{N} H_{jm} |v_m| \tag{9}$$

In the above derivation , the fact that all entries of hyperlink matix are non negative i.e. $H_{jm} \geq 0$ is used . For any column of hyperlink matrix H the condition $\sum_{j=1}^{N} H_{jm} = 1$ holds , which when substituted in (9) results in

$$|\lambda| v_m \leq v_m \Longrightarrow |\lambda| \leq 1 \tag{10}$$

□

Theorem 2.3. *If the hyperlink matrix H satisfies the following properties*

- *There is exactly one eigen value such that* $|\lambda| = 1$
- *The eigenspace associated the eigenvalue having* $|\lambda| = 1$ *is of dimension* 1
- *The hyperlink matrix H representing the web is diagonalizable ,i.e. its eigen vectors are independent and form a basis*

then the following holds:

1. *There exists a unique vector S s.t. the entry* $S_i = P(X_n = i)$ *indicates the page rank value of page i, which satisfies the following relations :*

$$S_i \geq 0 \tag{11}$$

$$S_i = \sum_{j=1}^{N} H_{ij} S_j \tag{12}$$

$$\sum_{i=1}^{N} S_i = 1 \tag{13}$$

The vector S is called the stationary regime of the Markov chain representing the activity of web surfing.

2. *The distribution of the probabilities($P(X_n = i)$) i.e. the final page rank values will converge to the stationary regime S as $n \to \infty$, irrespective of the initial setting of the page rank values i.e. p^0 (where $\sum_{i=1}^{N} p_i^0$)*

Considering the hyperlink model of Figure 1, the matrix H has an eigenvalue 1 with multiplicity 1 and 5 other eigenvalues, with 4 eigenvalues being complex and 1 real.The eigenvector associated with eigenvalue 1 is given by $[0.7306\ 0.4870\ 0.2740\ 0.0304\ 0.0913\ 0.3805]^t$, which is normalized by dividing by the sum of entries to obtain the stationary regime S as

$$S = \begin{bmatrix} 0.3664 \\ 0.2443 \\ 0.1374 \\ 0.0153 \\ 0.0458 \\ 0.1908 \end{bmatrix} \tag{14}$$

After sufficient number of traversals along the graph, the web surfer will visit the page A most often. Thus the ranking of pages is given by : A, B, C, E, F, D. The above result for the stationary regime can be verified by taking two arbitary initial probability distriutions i.e. $p_1^0 = [0.5\ 0.25\ 0\ 0\ 0.25\ 0]$ and $p_1^0 = [0.333\ 0.333\ 0.333\ 0\ 0\ 0]$, and arriving at the same result for stationary regime given by (14).

3 Teleportation model

The teleportation model is based on the assumption that the random surfer , after following the web structure in executing random walk , makes an arbitary jump to another page not connected directly to the current page. For the random jump the N webpages(here in Figure 1 N=6) have equal probability of being accessed by jumps. The parameter which determines the jump is given by m and changes the hyperlink matrix H to a modified matrix called Google matrix G. The matrix G is given by this following relation , also known as *PageRank equation*:

$$G = (1 - m)H + \frac{m}{n} QQ^T \tag{15}$$

$\frac{1}{n}QQ^T$ is a square matrix of rank 1, where each entry is $\frac{1}{n}$. Since the entries of matrix H and $\frac{1}{n}QQ^T$ are positive the entries of the updated matrix G are also positive.For the hyperlink model of figure (1), considering the hyperlink matrix in(5) and setting $m = 0.15$(this parameter value was considered in the original algorithm algorithm),we have the Google matrix as:-

$$G = \begin{bmatrix} 0.0250 & 0.4500 & 0.3083 & 0.4500 & 0.0250 & 0.8750 \\ 0.4500 & 0.0250 & 0.3083 & 0.0250 & 0.3083 & 0.0250 \\ 0.0250 & 0.4500 & 0.0250 & 0.0250 & 0.3083 & 0.0250 \\ 0.0250 & 0.0250 & 0.0250 & 0.0250 & 0.3083 & 0.0250 \\ 0.0250 & 0.0250 & 0.3083 & 0.0250 & 0.0250 & 0.0250 \\ 0.4500 & 0.0250 & 0.0250 & 0.4500 & 0.0250 & 0.0250 \end{bmatrix} \tag{16}$$

The updating equation of the page rank values is changed to

$$x(k + 1) = Gx(k) \tag{17}$$

Since all the entries of the google matix G are greater than zero and the column entries sum to 1, then by applying a Markov chain based analysis (simlar to the case in section 2), the page rank vector is obtained as :

$$S = \begin{bmatrix} 0.3385 \\ 0.2267 \\ 0.1397 \\ 0.0433 \\ 0.0646 \\ 0.1873 \end{bmatrix} \tag{18}$$

In the practical case the large dimension of google matrix G makes computation of eigen vectors diifi-cult.Google usually maintains a catalog of 10 billion webpages , which can be considered as the number of rows in the google matrix G. The computation of the eigen vector is done in an iterative fashion using equation (17).Since the Google matrix G is not sparse , then in order to improve the computation process the sparse nature of hyperlink matrix H is utilized by rewriting the updating equation (17)as follows:

$$x(k+1) = Gx(k) = (1-m)Hx(k) + \frac{m}{n}Q; \tag{19}$$

Here Q is a $N \times 1$ vector with all its entries equal to 1.The convergence details will be considered in the next section .

4 Perron Frobenius Theorem

This section in general details the Perron Frobenius Theorem [5] for an irreducible , non negative , square matrix. A matrix is called non negative if all its entries are greater than or equal to zero.The idea of irreducibility of a matrix is considered here before stating the theorem of Perron Frobenius. A matrix is said to be irreducible if the following holds:-

- The matrix is not similar via a permutation to a block upper traingular matrix. It means that for a permutation matrix P (A permutation matrix has one entry 1 in each row and each column and rest are all zeros),the following relation holds for an irreducible matrix A.

$$P^T AP \neq \begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix} \tag{20}$$

- In terms of graph theory if the matrix represents the adjacency matrix of the directed graph and is irreducible , then the graph is strongly connected i.e. every vertex can be reached from other vertex.
- For every pair of indices i and j, there exists a natural number m such that such that for an irreducible matrix A we have $(A^m)_{ij} > 0$

The Perron Frobenius theorem is detailed below:

Theorem 4.1 (Perron Frobenius Theorem)**.** *For any irreducible,non-negative,square matrix A, the following statements hold*

1. *Matrix A has a real eigenvalue r called the Perron root such that $r \geq 0$ and $r = \rho(A)$.*
2. *The algebraic multiplicity of r is 1.*
3. *There exists a strictly positive eigenvector $\boldsymbol{x}$ for r.*
4. *There exists a unique vector $\boldsymbol{p}$ called the **Perron vector** defined by*

$$\boldsymbol{Ap=rp} \qquad \boldsymbol{p} \geq 0 \qquad \|\boldsymbol{p}\|_1 = \boldsymbol{1}$$

5. *The Collantz Wielandt formula determines the Perron root of the irreducible square matrix as follows:*

$$\boldsymbol{r} = max_{x \in N} f(x) \qquad \boldsymbol{f(x)} = min_{1 \leq i \leq n \ x_i \neq 0} \frac{[Ax]_i}{x_i} \tag{21}$$

4.1 Application in Page Rank Algorithm analysis

At first a primitive matrix is defined as follows:

1. A nonnegative irreducible square matrix is said to be primitive if it has no other eigen value greater than or equal to $r = \rho(A)$.
2. A nonnegative matrix is primitive if there exists a natural number b such that $A^b > 0$.

The matrix G as denoted by (16) has all positive entries i.e. $G_{ij} > 0$. For any positive integer m we have $(G^m)_{ij} > 0$.Hence the matrix G is irreducible in nature.Thus the Perron Frobenius theorem is applicable on Google matrix G.
Then there exists an eigenvalue r of algebraic multiplicity 1 such that $r = \rho(G)$. The matrix G has the special property that $\sum_{i=1}^{N} G_{ij} = 1$, due to which an analysis similar to hyperlink matrix can be done to arrive at the result that $r = \rho(G) = 1$. There also exists a unique eigen vector p coressponding to $r = 1$ such that all its entries are positive and sum up to 1. For the matrix G given by (16) the eigenvector coressponding to eigenvalue 1 is given by

$$X = \begin{bmatrix} 0.7111 \\ 0.4763 \\ 0.2934 \\ 0.0909 \\ 0.1356 \\ 0.3934 \end{bmatrix} \tag{22}$$

Now for matrix G , if X is an eigenvector then cX is also an eigenvector , where c is any constant. Considering c to be the sum of entries in X i.e. 2.1007, then we obtain the unique Perron vector as:

$$p = \begin{bmatrix} 0.3385 \\ 0.2267 \\ 0.1397 \\ 0.0433 \\ 0.0646 \\ 0.1873 \end{bmatrix} \tag{23}$$

In general it is difficult to calculate eigen vectors by hand. So the popular procedure for calculating eigen vector is the **power method**, where an initial solution is picked and multiplied by G until the convergence is reached. The matrix G has all positive entries , which means that it is primitive according to the 2nd definition given at the start of this section.Thus according to definition 1, the magnitudes of all other eigen values are less than 1. Thus we have the following relation

$$\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq |\lambda_4| \geq |\lambda_5| \geq |\lambda_6| \tag{24}$$

Considering v_j to be the eigen vector coressponding to λ_j, then the initial solution $x(0)$ can be expressed as $x(0) = k_1v_1 + k_2v_2 + k_3v_3 + k_4v_4 + k_5v_5 + k_6v_6$. By application of power iteration procedure we have

$$\begin{aligned}
&x(1) = Gx(0) = k_1Gv_1 + k_2Gv_2 + \cdots + k_6Gv_6 = k_1\lambda_1v_1 + k_2\lambda_2v_2 + \cdots + k_6\lambda_6v_6 \\
&x(2) = Gx(1) = k_1\lambda_1^2v_1 + k_2\lambda_2^2v_2 + \cdots + k_6\lambda_6^2v_6 \\
&\vdots \\
&x(6) = Gx(5) = k_1\lambda_1^6v_1 + k_2\lambda_2^6v_2 + \cdots + k_6\lambda_6^6v_6
\end{aligned}$$

For $j \neq 1$, we have $\lambda_j < 1$, so λ_j^6 tends to 0 .Thus the result of power method converges to k_1v_1, i.e. the eigenvector coresponding to eigen value=1. This eigen vector, when suitably scaled gives the page rank values for different pages.
The convergence of the power method depends on the largest and the second largest eigen values given by $\lambda_1(G)$ and $\lambda_2(G)$ respectively. It depends on the ratio of the two eigen values $\frac{\lambda_2(G)}{\lambda_1(G)}$.Considering the parameter m in section 3 then $\lambda_2(G)$ is given as:

$$|\lambda_2(G)| \leq 1 - m \tag{25}$$

When the value of m becomes large then the convergence is faster but the effect of link structure i.e. hyperlink matrix will be reduced. So a compromise is made by considering $m = 0.15$.

5 Banach fixed point theorem

In this section the Banach fixed theorem is discussed in brief along with its application in the page ranking problem.Before discussing the Banach fixed point theorem, the ideas of metric space [6], normed linear space especially Banach space and contraction mapping are considered.

5.1 Metric space

A metric space is basically a set X with a metric defined on it.The metric associates with any pair of elements in set X, a distance d. Formally a metric space is defined as a pair (X, d) where X is the set and d is a metric on X i.e. a function defined on $X \times X$ such that for all $x, y, z \in X$ we have the following properties:

- d is real valued,finite and non-negative.
- $d(x, y) = 0$ if and only if x=y.
- $d(x, y) = d(y, x)$
- $d(x, y) \leqq d(x, z) + d(z, y)$

5.2 Normed linear space

Normed linear space can be considered as a case of metric space ,where the set X(as considered in the definition of metric space) is a vector space and the metric is given by means of a norm. A linear space χ consisting of vectors is a normed linear space if for each vector $x \in \chi$, we have a real valued norm $\|x\|$, which satisfies the following properties:

- $\|x\| \geq 0$ for all $x \in \chi$ with $\|x\| = 0$ if and only if $x = 0$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \chi$
- $\|\alpha x\| = |\alpha|\|x\|$ for all $\alpha \in R$ and $x \in \chi$

5.3 Banach Space

In order to describe Banach space the idea of Cauchy sequence is considered.
Cauchy Sequence: A sequence $x_k \in \chi$ is said to be Cauchy sequence if $\|x_k - x_m\| \to 0$ as $k, m \to \infty$.
Using the above idea of Cauchy sequence the Banach space is defined as follows:
Banach space: A normed linear space χ is complete if every Cauchy sequence in χ converges to a vector in χ. Banach space can be considered as a normed space which is a complete metric space.

5.4 Contraction Mapping

If S is a closed subset of a Banach space χ and g be the mapping that maps S to S.Then the mapping g is said to be contraction mapping if the following holds:

$$dist(g(x), g(y)) \leq \rho \; dist(x, y) \tag{26}$$

Here $0 \leq \rho < 1$.

5.5 Description of Banach fixed point theorem

The Banach fixed point theorem is stated below:

Theorem 5.1. *Consider a metric space $S = (S, d)$ where $S \neq \emptyset$.Suppose that S is complete and $g : S \to S$ is a contraction on S , then S has precisely one fixed point satisfying $x^* = g(x^*)$.The fixed point x^* can be obtained from any initial vector in S by method of successive approximation.*

5.5.1 Analysis of Page Rank Algorithm

In order to show that Banach's fixed point theorem [7] is applicable it is necessary to consider the set ω (consisting of probability vectors). The details of the set ω are given below:

$$\omega = \{x \in R^N | x \geq 0, 1^t x = 1\} \tag{27}$$

Lemma 5.2. *Let Y be the column stochastic $N \times N$ matrix. Then Y causes contraction for the 1-norm. For each $x \in \omega$, $Yx \in \omega$*

Proof. Consider $x = [x_1, x_2, x_3, \cdots x_N]^t \in R^N$, then the vector $|x|$ is given by $|x| = [|x_1|, |x_2|, |x_3|, \cdots |x_N|]^t$. Then the l_1 norm of x is given by

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_N| \tag{28}$$

Now for each entry of Yx is given by $(Yx)_i = Y_{i1}x_1 + Y_{i2}x_2 + \cdots + Y_{iN}x_N$. Then the following relations hold :-

$$\begin{aligned} |(Yx)_i| &= |Y_{i1}x_1 + Y_{i2}x_2 + \cdots + Y_{iN}x_N| \\ &\leq |Y_{i1}x_1| + |Y_{i2}x_2| + \cdots |Y_{iN}x_N| \\ &= Y_{i1}|x_1| + Y_{i2}|x_2| + \cdots Y_{iN}|x_N| \\ &= Y(|x|)_i \end{aligned} \tag{29}$$

For all $i \in N$ we have $|Yx| \leq Y|x|$. Thus we have another relation relating to l_1 norm , which is given as follows:

$$\begin{aligned} \|Yx\|_1 = 1^t|Yx| \leq 1^tY|x| = \|x\|_1 \\ \implies \|Yx\|_1 \leq \|x\|_1 \end{aligned} \tag{30}$$

This relation shows that Y is the contraction for 1 norm.Also for $x \in \omega$ and $Y \geq 0$ $Yx \geq 0$, thus satisfying one property of the set ω. Also $1^tYx = (1^tY)x = 1^tx = 1$.Thus the set ω is invariant under the action of a column stochastic matrix . Then the space ω represents a Banach space i.e. complete space. □

If $x, y \in \omega$ and the matix G is given by (15) then the following holds

$$\|Gx - Gy\|_1 = \|(1-m)Hx - (1-m)Hy\|_1 = (1-m)\|H(x-y)\|_1 \leq (1-m)\|x-y\|_1 \tag{31}$$

Here $1 - m \in [0, 1)$ which is similar to the parameter ρ in the definition of contraction mapping.Thus Banach fixed point theorem is applicable and by defining a sequence $x_{k+1} = Gx_k$, a fixed point x^* can be found out using the relation $x^* = Lim_{n\to\infty}G^nx_0$. The entries of the fixed point vector x^* give the page rank values of the different pages .This process of computing the final Pagerank vector is better because it does not employ computation of eigen values and eigen vectors.

6 Aggregation based Page Ranking system

This section includes an approximated technique of page rank computation [2] where the group structure of the web is taken into account , i.e. the page rank values are computed for different groups in the web followed by the distribution of the page rank values among the members. The following subsections highlight the ides of web aggregation and computation of approximated page rank value.

6.1 Web aggregation

Tge sparse nature of the web can be attributed to the existence of many intrahost links, due to which the linking of the pages are done within same organizations, departments and universities. For the purpose of depicting web aggregation the following diagrams are considered:
The steps that are followed in the page rank computation procedure on the basis of web aggregation are given as:

1. *Grouping stage:* First groups are searched in the web.
2. *Global step:* The Page rank is computed for each group separately.
3. *Local step:* The total page rank value is distributed among its group members.

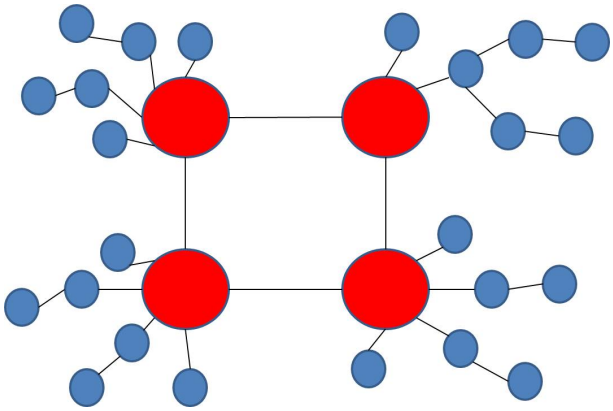

Figure 3: **Web graph (Red nodes indicate servers and the blue nodes indicate nodes con .**

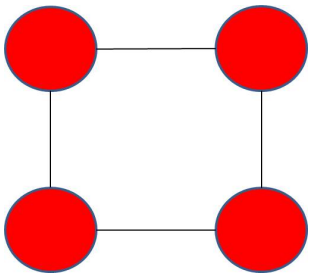

Figure 4: **Aggregated graph .**

The server level performs the grouping operation and then data is exchanged among the groups.In the local step, each group is responsible for carrying out its own operation. The above grouping approach can be performed in a decentralised fashion with the purpose of computing page rank values efficiently. Web structure is in general sparse. The sparsity of the web can be effectively captured using grouping operation. The node parameter that actually determines the sparse nature of the web is given by the following expression:

$$\delta_i = \frac{number\ of\ outgoing\ links\ from\ node\ i\ to\ other\ groups}{total\ number\ of\ outgoing\ links\ from\ node\ i} \tag{32}$$

Here the above mentioned parameter is defined for each node i.e. webpage and smaller value of the parameter indicates lesser number of outgoing links to other groups resulting in sparse structure.
Each node parameter for a particular node is bounded by $\delta_i \leq \delta$, where $\delta \in (0, 1)$. Those pages for which the number of external links to other groups are large are considered as single groups. For such pages the above mentioned node parameters always become 1 and the bound is not considered.
For the purpose of web aggregation ng is considered as the number of groups and ng_1 is the number of single groups. For group i the number of member pages is given by n_i. In the Page rank vector considered the first n_1 elements are for pages in group 1 and the next n_2 elements are for group 2 and so on.

6.2 Approximated Page Rank via Aggregation

The group value of a group i is given as the sum of the page rank values of the members of the group i.A coordinate transformation is performed as follows:-

$$\tilde{x}^* = \begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} x^* \tag{33}$$

Here the ith entry of $\tilde{x}_1^*$ is the group value of the ith group and $\tilde{x}_2^*$'s ith entry determines the difference between a page value and average value of the group members.The vector $\tilde{x}_1^*$ is the aggregated PageRank and x^* is the original page rank vector. The coordinate transformation can be properly depicted using the following figure:
The matrix W is detailed as below:

$$W = [W_1^T W_2^T]^T \tag{34}$$

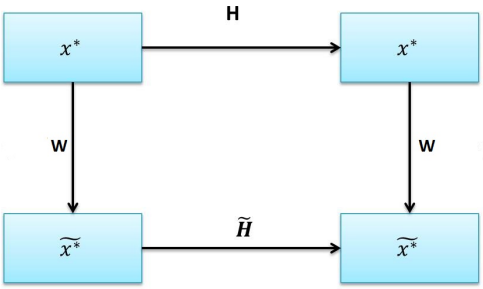

Figure 5: **Coordinate transformation** .

where W_1 and W_2 are block diagonal matrices , containing ng and $ng - ng_1$ blocks respectively.Details are given as:

$$W_1 = blockdiagonal(1_{n_i}^T) \in R^{r \times N} \tag{35}$$

$$W_2 = blockdiagonal([I_{n_i - 1} 0] - \frac{1}{n_i} 1_{n_i - 1} 1_{n_i}^T) \tag{36}$$

The matrix $\tilde{H}$ can be written as:

$$\tilde{H} = WHW^{-1} = \begin{bmatrix} \tilde{H_{11}} & \tilde{H_{12}} \\ \tilde{H_{21}} & \tilde{H_{22}} \end{bmatrix} \tag{37}$$

The page rank equation is modified as follows:

$$\begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} = (1-m) \begin{bmatrix} \tilde{H_{11}} & \tilde{H_{12}} \\ \tilde{H_{21}} & \tilde{H_{22}} \end{bmatrix} \begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} + \frac{m}{n} \begin{bmatrix} u \\ 0 \end{bmatrix} \tag{38}$$

Here $u = W_1 1_n = [n_1, n_2, n_3, \cdots n_{ng}]^T$. The submatrices of $\tilde{H}$ i.e. $\tilde{H_{11}}, \tilde{H_{12}}, \tilde{H_{21}}, \tilde{H_{22}}$ satisfy certain properties:

- $\tilde{H_{11}}$ is a stochastic matrix i.e. all its entries are positive and column entries sum upto 1.
- The matrix H can be broken down into three different matrices $H_{int}, H_{ext1}, H_{ext2}$, where H_{int} determines the internal links among the groups, H_{ext1} contains the column of H coressponding to single groups and H_{ext2} contains the remaining columns of the non single groups.This matrix H_{ext2} has only small entries outside the diagonal blocks and contribute to the submatrices $\tilde{H_{12}}$ and $\tilde{H_{22}}$. The entries of the matrix $\tilde{H_{12}}$ are small in magnitude .

By considering the above properties the approximated page rank is calculated by triangonalization of matrix $\tilde{H}$ resulting in the following equation:

$$\begin{bmatrix} \tilde{x}_1^{'} \\ \tilde{x}_2^{'} \end{bmatrix} = (1-m) \begin{bmatrix} \tilde{H_{11}} & 0 \\ \tilde{H_{21}} & \tilde{H_{22}^{'}} \end{bmatrix} \begin{bmatrix} \tilde{x}_1^{'} \\ \tilde{x}_2^{'} \end{bmatrix} + \frac{m}{n} \begin{bmatrix} u \\ 0 \end{bmatrix} \tag{39}$$

This approximated page rank computation is done in the changed coordinate system .In order to obtain the approximated page rank in the original coordinate system i.e. $x^{'}$ the following relation is followed:

$$x^{'} = W^{-1} \tilde{x}^{'} \tag{40}$$

6.3 Approximated Page Rank computation

The steps for computation of approximated page rank values, starting from an initial stochastic vector $\tilde{x}(0)$ are detailed as follows:-

- Using the first row of (39) we obtain the first state as:-

$$\tilde{x}_1(k+1) = (1-m)\tilde{H_{11}}\tilde{x}_1(k) + \frac{m}{n}u \tag{41}$$

- From the second row the second state $\tilde{x}_2(k)$ is obtained as

$$\tilde{x}_2(k) = (1-m)[1 - (1-m)\tilde{H_{22}^{'}}]^{-1}\tilde{H_{21}}\tilde{x_1}(k) \tag{42}$$

- Then the transformation is done to the original coordinate system as

$$x(k) = W^{-1}\tilde{x}(k) \tag{43}$$

7 Application of Page Ranking system: Ranking of Journals

The idea of page ranking can be extended to the computation of impact factors of journals [2]. When a user reads a journal, he also accesses journals which are cited in the current one resulting in a random walk across different journals, which is similar to the random walk of the web surfer. Since the act of a journal getting cited is probabilistic in nature , a particular metric similar to the page rank value, called $Eigen\ factor\ score(EF)$ is considered.
A matrix called cross citation matrix C similar to the hyperlink matrix H for the page ranking system is defined as follows:

$$C_{ij} = \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \tag{44}$$

Here each entry a_{ij} is defined as

$$a_{ij} = \begin{cases} \begin{array}{c} number\ of\ citations\ in\ a\ particular\ year \\ from\ journal\ j\ to\ articles \\ published\ in\ journal\ i\ in\ a\ fixed\ span\ of\ few\ years \end{array} & i \neq j \\ 0 & i = j \end{cases} \tag{45}$$

When some journals dont cite any other journals then their columns will only consist of zero entries. To resolve this problem the cross citation matrix C is redefined by considering the idea of article vector $v = [v_1, v_2, v_3, \cdot \cdot v_n]^T$, where we have

$$v_j = \frac{number\ of\ articles\ published\ in\ journal\ j\ in\ a\ particular\ time\ span}{number\ of\ articles\ published\ by\ all\ journals\ in\ that\ particular\ time\ span} \tag{46}$$

The zero columns of cross citation matrix C are relpaced by the entries of the article vector v to obtain an updated matrix $\tilde{C}$.The resulting cross citation matrix will be a stochastic matrix and the eigen factor equation similar to (15) will be given as:

$$M = (1 - m)\tilde{C} + mv1^T \tag{47}$$

Here m is fixed to 0.15 and 1 is a column vector consisting of all ones. The analysis performed for the PageRank problem will hold, thus resulting in an eigenvector x^* coressponding to the largest eigen value of matrix M. This eigen vector x^* is called the journal influence vector, resulting in different weights on citation values.

8 Conclusions

This article details the working of the Page ranking system by considering the simple random walk model and the updated system which is based on random jumps from one web page to another. The description is followed by suitable mathematical analysis , where the idea of Markov chain has been considered and the theorems of Perron Frobenius for irreducible , positive matrices and Banach for contracion mapping have been invoked.Further , a new system of approximate page ranking based on aggregration has been discussed followed by an example of page ranking in computing importance of the journals.

References

[1] Sergey Brin, Larry Page, *"The anatomy of a large scale hypertextual Web search engine"*Comput.Netw.ISDN Syst.,vol.30,nos.1-7,pp.107-117,1998

[2] Hideaki Ishii, Roberto Tempo *"The PageRank Problem,Multiagent Consensus and Web Aggregation A Systems and Control Viewpoint "*IEEE Contol Systems Magazine, June 2014

[3] Google webpage (2014) http://www.google.co.in/insidesearch/howsearchworks/thestory/

[4] Christiane Rousseau, Yvan Saint Aubin *"Mathematics and Technology "*Springer Undergraduate texts,2008

[5] Carl D Meyer *"Matrix Analysis and Applied Linear Algebra "*SIAM,2000

[6] Erwin Kreyszig *"Introductory Functional Analysis with Applications "*John Wiley and Sons,1978

[7] Markus Haase *"Functional Analysis: An Elementary Introduction "*Volume 156, American Mathematical Society