THEORY OF SEARCH ENGINES

K. K. NAMBIAR

ABSTRACT. Four different stochastic matrices, useful for ranking the pages of the Web are defined. The theory is illustrated with examples.

Keywords—Search engines, Page rank, Web.

CONTENTS

1. INTRODUCTION

Granted that the Web contains an incredible amount of information, it should be admitted that the problem of finding the relevant one of interest to you is by no means easy. Recently, there has been attempts to circumvent this difficulty by ranking the pages of the web according to their importance, and list only those pages of high rank containing the required information. The popular Google employs this method [2] in their search engine, and much of their success is attributed to their method of ranking the pages in the Web. The purpose of this paper is to delineate the theory that can be used in ranking the pages and hyperlinks of the Web.

It is obvious that the Web can be considered as a directed graph with pages as nodes and hyperlinks as directed edges. In the Web graph (webgraph), we will say that two nodes are connected, if there are directed paths both ways between the two nodes. Clearly, connectedness is an equivalence relation and hence it splits the set of nodes of the webgraph into disjoint subsets. The subgraph induced by such a subset, we will call a community graph or just a community. If we coalesce all the nodes of a community and consider it as a single node (community node), the webgraph reduces to an acyclic graph, in which no two community nodes can have directed paths both ways. This graph we will call the reduced webgraph and represent the indegree of each community node by e_k. We define the *exponential*

Date: January 8, 2001.

mean of $\{e_k\}$ as

$$\nu = \prod_k {e_k}^{c_k}$$

where $c_k = e_k/E$ and $E = \sum_k e_k$.

Since the indegree of a community node gives an indication of its importance in the Web, it will not be unreasonable for us to give special importance to a community, whose indegree is not less than the exponential mean ν.

Consider a Web surfer starting from an arbitrary node in a community and surfing within the community in all possible ways. The number of different paths the surfer can take can be infinite and at no time will the surfer be faced with the situation when he does not have a hyperlink available to click. We define the *freedom* in a community as the logarithm (base 2) of the increase in the number of different paths per mouse click. Specifically, we define freedom as

$$C = \lim_{N\to\infty} \frac{\log_2 P_N}{N}$$

where P_N is the number of different paths possible with N clicks. We define *activity* in the community as $A = 2^C$. It is known that the activity of the community will not depend on the initial starting point of the surfer.

2. RANKING OF PAGES

We will show that the community graph can be used to calculate the activity in the community. The graph can be used also to rank the pages and hyperlinks of the community. The crucial concepts we will use to calculate these are the Perron-Frobenius theorem and Shannon's theory of communication [1, 3].

The communities in the Web consist of millions of pages and hyperlinks, yet it is a finite graph and hence we can study their properties by analyzing graphs with a few number of nodes, which is what we will do here.

The usual way to analyze a matrix is through its characteristic equation and eigenvectors, here we do the same thing, except that we start off with a slight variation, we call, intrinsic equation. Our first job is to characterize a community by a matrix. An example of our notation is,

$$\mathbf{N}(s) = \begin{bmatrix} 3s & s & 2s \\ 3s & 5s & 6s \\ s & s & 4s \end{bmatrix}$$

which characterizes a community with three pages. The (i, j) element specifies the number of hyperlinks from page i to page j. For example, the $(2, 1)$ element $3s$ says that there are three hyperlinks from page 2 to page 1, and similar explanation goes for other elements also. The purpose of the variable s will be clear from the next example.

The above notation can be extended slightly to include hyperlinks that go through intermediate nodes. Consider the matrix,

$$\mathbf{N}(s) = \begin{bmatrix} 0 & s^2 + s^4 \\ s^3 + s^6 & s^2 + s^4 \end{bmatrix}.$$

Here, the $(2,1)$ element $s^3 + s^6$ says that there are two ways to reach page 1 from page 2, one by taking three hops, another by taking six hops, through pages of no interest to us.

We will use these two matrices as examples to illustrate our method of ranking pages and links. We need some definitions and notations to proceed any further.

$\mathbf{N}(s)$: The matrix which characterizes the community, as in the two examples above.

I: A unit matrix of appropriate dimension.

$\mathbf{I} - \mathbf{N}(s)$: The intrinsic matrix of the community. The connection between the usual characteristic matrix and the intrinsic matrix can be recognized, if we put $s = \lambda^{-1}$.

$\det\{\mathbf{I} - \mathbf{N}(s)\} = 0$: The intrinsic equation of the community.

μ: The smallest positive root of the intrinsic equation. Perron-Frobenius theorem assures us that there will always be a smallest positive root.

C: A measure of the freedom in the network, as defined earlier. Shannon's theory of communication tells us that $C = -\log_2 \mu$.

A: A measure of the activity in the network, as defined earlier. Shannon's theory of communication tells us that $A = \mu^{-1}$.

p: The intrinsic column vector of $\mathbf{N}(s)$. The intrinsic column vector is defined as the solution of $\mathbf{N}(\mu)\mathbf{p} = \mathbf{p}$, with positive elements, the sum of the elements being unity. Perron-Frobenius theorem assures us the existence of such a $\mathbf{p}$. We will call this matrix, customer-ranking matrix or *c-matrix*, since the magnitude of its elements gives the prominence of the page as a customer.

q: The intrinsic row vector of $\mathbf{N}(s)$. The intrinsic row vector is defined as the solution of $\mathbf{q}\mathbf{N}(\mu) = \mathbf{q}$, with positive elements, the sum of the elements being unity. Perron-Frobenius theorem assures us the existence of such a $\mathbf{q}$. We will call this matrix, vendor-ranking matrix or *v-matrix*, since the magnitude of the elements of the matrix gives the prominence of the page as a vendor.

r: A row matrix that can be used to rank the pages of the community. It is defined as $(\mathbf{qp})^{-1}[p_j q_j]$. We will call this matrix, *r-matrix*.

P: A square matrix that can be used to rank pairs of pages in the community. It is defined as $[p_i q_j]$. We will call this matrix, link-ranking matrix or *l-matrix*, since the magnitude of the elements of the matrix gives the prominence of links between pairs of pages. The diagonal elements here give the ranking of the corresponding pages.

e_k: The indegree of the k^{th} community node in the Web.

E: The total number of edges in the reduced graph, $\sum_k e_k$, as defined earlier.

c_k: The fractional degree of the k^{th} community node, e_k/E, as defined earlier.

3. TWO EXAMPLES

Example 1.

$$\mathbf{N}(s) = \begin{bmatrix} 3s & s & 2s \\ 3s & 5s & 6s \\ s & s & 4s \end{bmatrix}$$

The required computations here are well-known and hence we omit all details. $\det\{\mathbf{I} - \mathbf{N}(s)\} = 0$ gives the roots as 0.5, 0.5, and 0.125.

$$\begin{aligned} \mu &= 0.125 \\ C &= -\log_2 \mu = 3 \\ A &= \mu^{-1} = 8 \end{aligned}$$

$$\mathbf{p} = \begin{bmatrix} 0.2 \\ 0.6 \\ 0.2 \end{bmatrix}$$

$$\mathbf{q} = \begin{bmatrix} 0.25 & 0.25 & 0.50 \end{bmatrix}$$

$$\mathbf{r} = \frac{10}{3} \begin{bmatrix} 0.05 & 0.15 & 0.10 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 0.05 & 0.05 & 0.10 \\ 0.15 & 0.15 & 0.30 \\ 0.05 & 0.05 & 0.10 \end{bmatrix}$$

Example 2.

$$\mathbf{N}(s) = \begin{bmatrix} 0 & s^2 + s^4 \\ s^3 + s^6 & s^2 + s^4 \end{bmatrix}.$$

$$\begin{aligned} \mu &= 0.688278 \\ C &= -\log_2 \mu = 0.538937 \\ A &= \mu^{-1} = 1.4529 \end{aligned}$$

$$\mathbf{p} = \begin{bmatrix} 0.411122 \\ 0.588878 \end{bmatrix}$$

$$\mathbf{q} = \begin{bmatrix} 0.301855 & 0.698145 \end{bmatrix}$$

$$\mathbf{r} = \begin{bmatrix} 0.231865 & 0.768135 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 0.124099 & 0.287023 \\ 0.177756 & 0.411122 \end{bmatrix}$$

4. CONCLUSION

Any of the four matrices $\mathbf{p}, \mathbf{q}, \mathbf{r}$, or $\mathbf{P}$ can be used to rank the pages of a community. For example, if we are interested in pages considered as vendors, we will compare the elements of the matrix $\mathbf{q}$ and list those pages corresponding to the larger elements of $\mathbf{q}$. The ranking of the vendor pages of the whole webgraph will require the comparison of the elements of the entire set of matrices $\{c_k\mathbf{q}_k\}$ where $\mathbf{q}_k$ is the v-matrix of the k^{th} community. From the published literature, it seems that a variation of this method has been used by Google in their search engine.

As an aside, we may state that the matrix $\mathbf{N}(s)$ in Example 2 can be considered as a model for the telegraph channel, as defined by Shannon [3]. This shows that there is considerable overlap between the theories of search engines and communication channels. For example, the freedom in the community in Example 2 is the same as the capacity of the telegraph channel.

REFERENCES

1. R. B. Bapat and T. E. S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, Cambrige, UK, 1997.
2. Sergey Brin and Laurence Page, *The Anatomy of a Large Scale Hypertextual Web Search Engine*, Proceedings of the Seventh International WWW Conference (Brisbane, Australia), April 14-18, 1998.
3. C. E. Shannon, *The Mathematical Theory of Communication*, University of Illinois Press, Chicago, 1963.

FORMERLY, JAWAHARLAL NEHRU UNIVERSITY, NEW DELHI, 110067, INDIA
Current address: 1812 Rockybranch Pass, Marietta, Georgia, 30066-8015
E-mail address: nambiar@mediaone.net