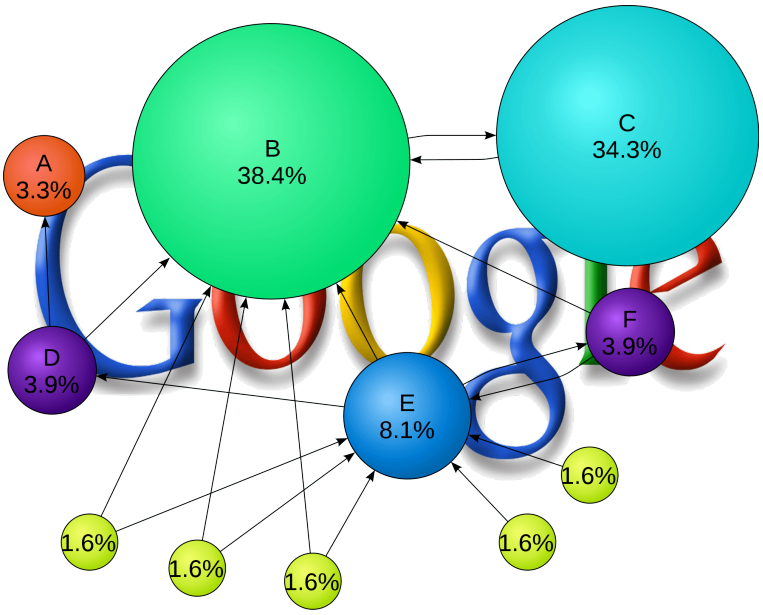
THE EIGENSPACE OF THE GOOGLE PAGERANK ALGORITHM (AN APPLICATION OF EIGENVALUES AND EIGENVECTORS)



Goal: To use eigenvalues and eigenvectors of a matrix for ranking the web-pages using the PageRank algorithm.

Initial Instruction :

- Open a new Matlab script and save it as name_PageRank.m.
- **Matlab commands used:** `eye, eig, diag, abs, fprintf, for ...end, sort, inv, det, norm, while ...end, if-elseif-end`

1. Conceptual Overview

The goal of this project is to use concepts from linear algebra to describe Google's PageRank algorithm. The goal of PageRank is to determine how "important" a certain web-page is. For example, Wikipedia is a more important web-page than stickers.com. But how can one quantify this importance? The idea is based on the following premises:

- The importance of web-page A is measured by the likelihood that an arbitrary web-surfer S will visit page A.
- The most likely way for web-surfer S to reach web-page A is by clicking on a link to A from web-page B (as opposed to randomly choosing web-page A out of the billions of potential web-pages).

Image refrence: https://www.sycosure.com/blog/google-pagerank-check/

- Web-surfer S is more likely to click on the link to A from B (rather than on a link to some other web-page C from B) if there are more web-link instances of A in B than C.
- In order to click on a link from B to A, web-surfer S must already be on web-page B.

In other words, **important pages are linked to lots of other important pages. However, if a web-page links to many other web-pages, the value of each link is watered down.**

As an example, consider the network of pages shown in the Figure 1 below. Web-pages A,B, C, and D are the least important pages because no other page connects to them. Next we consider web-pages E and F. E and F can be accessed from three other pages (A, B, and C and A, D, and E respectively). However, one of the the pages linking to F (page E) is important whereas none of the pages linking to E are important. Therefore, page F is more important than page E. Furthermore, page A contributes less to the importance of pages E and F than pages B, C, and D do since it links to two pages (whereas pages B, C, and D each only link to one page). However, these extra links do not affect the ranking of A; but only the rank of the pages linked to by A.

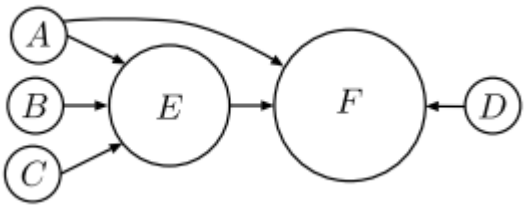


FIGURE 1. Illustration

2. MATHEMATICAL MODEL

To make this more quantitative, let R_A be the ranking of page A. Let the links between pages be defined as follows:

$$l_{ij} = \begin{cases} 1, & \text{if there exists a link from i to j,} \\ 0, & \text{if there does not exist a link from i to j.} \end{cases}$$

We assume that $l_{ii} = 0$ for all pages i so that a page can't link to itself. Let

$$n_i = \sum_{j \in \text{page}} l_{ij} \qquad \text{where page} = \{A, B, C, D, E, F\}$$

be the total number of links from page i to other pages. Then the ranking of a page equals the sum of the number of links from other pages weighted by the ranking of the

other pages and by the total number of links from the other pages.

$$R_i = \sum_{j \in \text{page}} \frac{l_{ji}}{n_j} R_j.$$

So, in the network shown in Figure 1, $R_E = \frac{1}{2}R_A + R_B + R_C$. We assume that the ranks sum to one, that is,

$$R_A + R_B + R_C + R_D + R_E + R_F = 1$$

3. Questions

Your task is to describe the PageRank model using concepts from linear algebra. For the purpose of this project, consider the network shown in the Figure 2 .

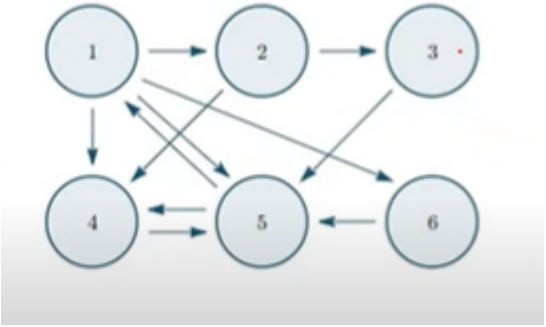


Figure 2. Network of Pages

Question 1. *Frame the given problem as an eigenvalue-eigenvector problem of the form $G\vec{v} = \lambda\vec{v}$ by clearly specifying the matrix G.*

Question 2. *Using `MATLAB`, determine all the eigenvalues and eigenvectors of the matrix G. Check whether the matrix G is diagnolizable or not.*

Question 3. *Write a `MATLAB` code to rank the given web-pages using the power method $\vec{v}_{k+1} = G.\vec{v}_k$ with initial guess $\vec{v}_0 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)^T$. Is the vector $\vec{v}_{k+1}$ stationary for some value of k? Also deduce the relationship between the eigenvectors and this stationary value.*

Question 4. *Now create your own web portal consisting of these six web-pages in which all of the web-pages share the hyperlinks of other page and rank them using the code developed for question (3). What difference do you find between the ranking of web-pages in questions (3) and (4).*

Question 5. *From the network shown in Figure 2, remove the link from page 4 to page 5 (assign $l_{45} = 0$). Will this effect the previous ranking of pages? Then construct a methodology which will help to rank the pages.*