Far East Journal of Mathematical Education

Published Online: May 2015
http://dx.doi.org/10.17654/FJMEFeb2015_055_068
Volume 14, Number 1, 2015, Pages 55-68 **ISSN: 0973-5631**

A SUPERINTRODUCTION TO GOOGLE MATRICES FOR UNDERGRADUATES

Kazuyuki Fujii and Hiroshi Oike

International College of Arts and Sciences
Yokohama City University
Yokohama, 236-0027, Japan
e-mail: fujii@yokohama-cu.ac.jp

Takado 85-5
Yamagata, 990-2464, Japan
e-mail: oike@tea.ocn.ne.jp

Abstract

In this paper, we consider so-called Google matrices and show that all eigenvalues (λ) of them have a fundamental property $|\lambda| \leq 1$. The stochastic eigenvector corresponding to $\lambda = 1$ called the PageRank vector plays a central role in the Google's software. We study it in detail and present some important problems.

The purpose of the paper is to make the heart of Google clearer for undergraduates.

1. Introduction

Google is one of important tools to analyze modern society. In this paper, we want to explain a secret of Google, which is "the heart of Google's software", to undergraduates.

Received: February 24, 2015; Accepted: March 16, 2015
Keywords and phrases: Google matrices, eigenvalue 1, PageRank vector, linear algebra.
Communicated by K. K. Azad

Although we are not experts of IT (information technology) the secret is clearly expressed in terms of linear algebra in mathematics. However, it is almost impossible to solve the linear algebra version explicitly, so we need some approximate method.

First, we give a fundamental lemma to understand a Google matrix (see the definition in the text) and present an important problem to define a realistic Google matrix (in our terminology). The problem is a challenging one for young researchers. For such a matrix we can use the power method to obtain the PageRank vector.

Second, we pick up an interesting example in [1] and calculate it thoroughly by use of MATHEMATICA. A good example and a thorough calculation help undergraduates to understand.

Last, we show an example which does not give the PageRank vector in terms of the power method with usual initial vector when H is not a realistic Google matrix. For this case we treat the power method with another initial vector and present a general problem.

We expect that undergraduates will cry out “I got Google!” after reading the paper.

2. Main Result

We introduce a Google matrix (realistic Google matrix) and study its key property.

We consider a collection of web pages with links (for example, a homepage and some homepages cited in it). See the figure in the next section (eight web pages with several links).

If a page has k links we give the equal weight $\frac{1}{k}$ to each link and construct a column vector consisting of these weights. See the figure once more. For example, since the page 4 links to the pages 2, 5 and 6 (three links) each weight is $\frac{1}{3}$. Therefore we obtain the column vector like

$$\text{page } 4 \rightarrow \begin{pmatrix} 0 \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \end{pmatrix} \begin{matrix} \\ < 2 \\ \\ \\ < 5 \\ < 6 \\ \\ \end{matrix}.$$

As a result, the collection of web pages gives a square matrix

$$H = (H_{ij}); \quad H_{ij} \geq 0, \ \sum_i H_{ij} = 1 \tag{2.1}$$

which we will call a *Google matrix*. Note that $H_{ii} = 0$ for all i (we prohibit the self-citation). From the definition it is a *sparse* matrix because the number of links starting from a webpage is in general small compared to the number of webpages.

If we set

$$J = (1, 1, ..., 1)^T,$$

where T is the transpose (of a vector or a matrix) then it is easy to see

$$H^T J = J \tag{2.2}$$

because row vectors of H^T are the transpose of column vectors of H like

$$\text{page } 4 \rightarrow \left(0, \frac{1}{3}, 0, 0, \frac{1}{3}, \frac{1}{3}, 0, 0\right).$$

From this we know that 1 is an eigenvalue of H^T. By the way, the eigenvalues of H are equal to those of H^T because

$$0 = |\lambda E - H| = |\lambda E - H^T|,$$

so we conclude that 1 is just an eigenvalue of H.

Therefore, we have the equation

$$HI = I, \tag{2.3}$$

where we assume that the eigenvector I is stochastic (the sum of all entries is 1). This I is called the *PageRank vector* and plays a central role in Google.

Now, we give a fundamental lemma to Google matrices:

Lemma. *Let* λ *be any eigenvalue of a Google matrix H. Then we have*

$$|\lambda| \leq 1. \tag{2.4}$$

The proof is easy and is derived from the Gerschgorin's (circle) theorem [2]. Note that the eigenvalues of H are equal to those of H^T and the sum of all entries of each row is 1 (see for example (3.2)). Namely,

$$\sum_{j=1}^{n} (H^T)_{ij} = \sum_{j=1}^{n} H_{ji} = 1 \text{ and } (H^T)_{ii} = H_{ii} = 0 \tag{2.5}$$

for all i and j.

We are in a position to state the Gerschgorin's theorem. Let $A = (a_{ij})$ be a $n \times n$ complex (real in our case) matrix, and we set

$$R_i = \sum_{j=1,\, j \neq i}^{n} |a_{ij}|$$

and

$$D(a_{ii};\, R_i) = \{z \in \mathbf{C} \,||\, z - a_{ii}| \leq R_i\}$$

for each i. This is a closed disc centered at a_{ii} with radius R_i called the *Gerschgorin's disc*.

Theorem (Gerschgorin)**.** *For any eigenvalue* λ *of A we have*

$$\lambda \in \bigcup_{i=1}^{n} D(a_{ii};\, R_i). \tag{2.6}$$

The proof is simple. Let us consider the equation

$$A\mathbf{x} = \lambda\mathbf{x} \quad (\mathbf{x} \neq \mathbf{0}) \tag{2.7}$$

and $|\, x_i \,|$ be the maximum

$$|\, x_i \,| = \max\{|\, x_1 \,|, |\, x_2 \,|, ..., |\, x_n \,|\} \neq 0 \Rightarrow \left| \frac{x_j}{x_i} \right| \leq 1.$$

From (2.7) we have

$$\sum_{j=1}^{n} a_{ij} x_j = \lambda x_i \Leftrightarrow \sum_{j=1,\, j\neq i}^{n} a_{ij} x_j = \lambda x_i - a_{ii} x_i = (\lambda - a_{ii}) x_i.$$

$x_i \neq 0$ gives

$$\lambda - a_{ii} = \sum_{j=1,\, j\neq i}^{n} a_{ij} \frac{x_j}{x_i}$$

and we have

$$|\, \lambda - a_{ii} \,| = \left| \sum_{j=1,\, j\neq i}^{n} a_{ij} \frac{x_j}{x_i} \right| \leq \sum_{j=1,\, j\neq i}^{n} \left| a_{ij} \frac{x_j}{x_i} \right| = \sum_{j=1,\, j\neq i}^{n} |\, a_{ij} \,| \left| \frac{x_j}{x_i} \right|$$

$$\leq \sum_{j=1,\, j\neq i}^{n} |\, a_{ij} \,| = R_i.$$

This means $\lambda \in D(a_{ii};\, R_i)$ for some i and completes the proof.

Finally, let us complete our lemma. In our case $H_{ij} \geq 0$, $H_{ii} = 0$ and $R_i = 1$ for all i and j, so these give the result

$$|\, \lambda \,| \leq 1$$

for any eigenvalue λ of H. This is indeed a fundamental property of Google matrices.

A comment is in order. The lemma must have been known. However, we could not find such a reference within our efforts.

Let us go ahead. In order to construct the eigenvector I in (2.3) a method called *the power method* is very convenient for a sparse matrix H of huge size. To calculate the characteristic polynomial is actually impossible.

The method is very simple, [1]. A sequence $\{I_n\}$ is defined recurrently by

$$I_n = HI_{n-1} \text{ and } I_0 = \mathbf{e_1}, \tag{2.8}$$

where the initial vector is $\mathbf{e_1} = (1, 0, ..., 0)^T$, which is usually standard. This is also rewritten as

$$I_n = H^n I_0 = H^n \mathbf{e_1}.$$

If $\{I_n\}$ converges to I, then we obtain the equation (2.3) like

$$HI = H(\lim_{n\to\infty} I_n) = \lim_{n\to\infty} HI_n = \lim_{n\to\infty} I_{n+1} = I.$$

In order that the power method works correctly some assumption on H is required. Namely,

- For a set of eigenvalues $\{\lambda_1 = 1, \lambda_2, ..., \lambda_n\}$ we assume

$$\lambda_1 = 1 > |\lambda_2| \geq \cdots \geq |\lambda_n|. \tag{2.9}$$

Note that 1 is a simple root. The assumption may be strong.

If a Google matrix H satisfies (2.9) we call H a *realistic* Google matrix. Now, let us present an important.

Problem. For a huge sparse matrix H propose a method to find or to estimate the second eigenvalue λ_2 without calculating the characteristic polynomial.

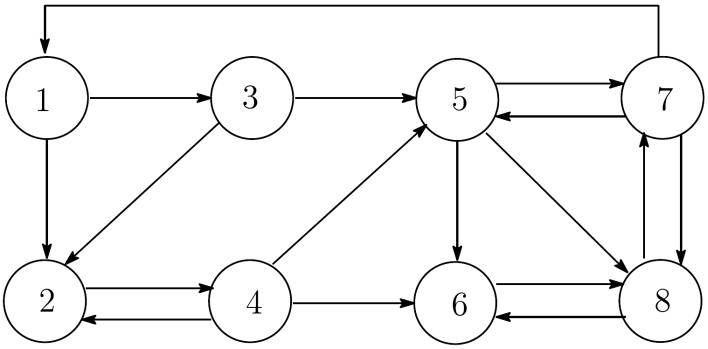
As far as we know such a method has not been given in mathematical physics or in quantum mechanics. This is a challenging problem for mathematical physicists.

3. Example

We consider an interesting example given in [1] and calculate it thoroughly by use of MATHEMATICA. A good example helps undergraduates to understand a model deeply.

In this section we need some results from linear algebra, so see for example [3] or [4] (we do not know a standard textbook of Linear Algebra in Europe or America or, etc.).

Example. A collection of web pages with links[1]



The Google matrix for this graph is given by

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{pmatrix} \tag{3.1}$$

[1]It is not easy for us to draw a (free) curve by use of the free soft WinTpic.

and its transpose is

$$H^T = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}. \tag{3.2}$$

If we define a stochastic vector

$$J = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)^T$$

it is easy to see

$$H^T J = J. \tag{3.3}$$

Let us study H from the mathematical view point by use of MATHEMATICA. The characteristic polynomial of H is given by

$$f(\lambda) = |\lambda E - H| = \begin{vmatrix} \lambda & 0 & 0 & 0 & 0 & 0 & -\frac{1}{3} & 0 \\ -\frac{1}{2} & \lambda & -\frac{1}{2} & -\frac{1}{3} & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & -\frac{1}{3} & \lambda & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & \lambda & 0 & -\frac{1}{2} \\ 0 & 0 & 0 & 0 & -\frac{1}{3} & 0 & \lambda & -\frac{1}{2} \\ 0 & 0 & 0 & 0 & -\frac{1}{3} & -1 & -\frac{1}{3} & \lambda \end{vmatrix}$$

$$= \lambda(\lambda - 1)\left(\lambda^6 + \lambda^5 - \frac{1}{9}\lambda^4 - \frac{1}{6}\lambda^3 + \frac{7}{108}\lambda^2 + \frac{11}{216}\lambda + \frac{1}{72}\right). \quad (3.4)$$

The exact solutions are $\{\lambda_1 = 1, \lambda_8 = 0\}$ and approximate ones (we round off a real number to five decimal places like $-0.87021\cdots = -0.8702$) are given by

$$\lambda_2 = -0.8702, \quad \lambda_3 = -0.5568,$$

$$\lambda_4 = 0.4251 - 0.2914i, \quad \lambda_5 = 0.4251 + 0.2914i,$$

$$\lambda_6 = -0.2116 - 0.2512i, \quad \lambda_7 = -0.2116 + 0.2512i.$$

From these we have

$$\lambda_1 = 1 > |\lambda_2| > |\lambda_3| > |\lambda_4| = |\lambda_5| > |\lambda_6| = |\lambda_7| > \lambda_8 = 0. \quad (3.5)$$

H becomes a realistic Google matrix from (2.9).

Moreover, the eigenvector for $\lambda_1 = 1$ is given by

$$\hat{I} = (24,\ 27,\ 12,\ 27,\ 39,\ 81,\ 72,\ 118)^T.$$

To check this (by hand) is not difficult and good exercise for undergraduates.

Since the sum of all entries of $\hat{I}$ is 400 the stochastic eigenvector (= the PageRank vector) I becomes

$$I = \begin{pmatrix} \frac{24}{400} \\ \frac{27}{400} \\ \frac{12}{400} \\ \frac{27}{400} \\ \frac{39}{400} \\ \frac{81}{400} \\ \frac{72}{400} \\ \frac{118}{400} \end{pmatrix} = \begin{pmatrix} 0.06 \\ 0.0675 \\ 0.03 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.18 \\ 0.295 \end{pmatrix}. \tag{3.6}$$

As a result, the ranking of webpages becomes

$$\text{page}\,8 > \text{page}\,6 > \text{page}\,7 > \text{page}\,5 > \text{page}\,2 = \text{page}\,4 > \text{page}\,1 > \text{page}\,3. \tag{3.7}$$

See the figure once more.

Here, let us show the power method to obtain the PageRank vector I, which is very useful if a realistic Google matrix is huge. A sequence $\{I_n\}$ is defined as

$$I_n = HI_{n-1} \text{ and } I_0 = (1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)^T$$

or

$$I_n = H^n I_0.$$

If the condition $|\lambda_2| < 1$ holds, then we have

$$\lim_{n\to\infty} I_n = I$$

because H can be diagonalized to be

$$H = S\text{diag}(1, \lambda_2, ..., \lambda_8)S^{-1} \Rightarrow H^n = S\text{diag}(1, \lambda_2^n, ..., \lambda_8^n)S^{-1}$$

with a matrix S consisting of eigenvectors. The speed of convergence depends on $|\lambda_2|$.

Let us list the calculation (rule: a real number is rounded off to five decimal places):

$$I_{40} = \begin{pmatrix} 0.0601 \\ 0.0675 \\ 0.0299 \\ 0.0676 \\ 0.0976 \\ 0.2022 \\ 0.1797 \\ 0.2954 \end{pmatrix}, \; I_{45} = \begin{pmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2024 \\ 0.1800 \\ 0.2951 \end{pmatrix}, \; I_{50} = \begin{pmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2024 \\ 0.1799 \\ 0.2951 \end{pmatrix}, \; I_{55} = \begin{pmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{pmatrix} \equiv I. \tag{3.8}$$

The result must be related to the powers of $|\lambda_2| = 0.87$ like

$$(0.87)^{40} = 0.0038, \;\; (0.87)^{45} = 0.0019, \;\; (0.87)^{50} = 0.0009,$$

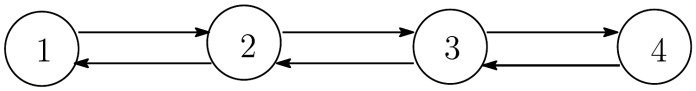
$$(0.87)^{55} = 0.0005. \tag{3.9}$$

Problem. Clarify a relation between I_n and $(0.87)^n$.

4. Counter Example

We show an example which does not give the PageRank vector in terms of the power method with usual initial vector $\mathbf{e}_1$ when H is not a realistic Google matrix.

Example. A collection of web pages with links



The Google matrix for this graph is given by

$$H = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{4.1}$$

The characteristic polynomial of H is given by

$$f(\lambda) = |\lambda E - H| = \lambda^4 - \frac{5}{4}\lambda^2 + \frac{1}{4} = (\lambda^2 - 1)\left(\lambda^2 - \frac{1}{4}\right) \tag{4.2}$$

and the solutions are

$$\lambda = \pm 1, \pm \frac{1}{2}. \tag{4.3}$$

Therefore, H is not a realistic Google matrix because of $\lambda = -1$. See (2.9) once more. For H it is easy to see that the PageRank vector is given by

$$I = \begin{pmatrix} \frac{1}{6} \\ \frac{2}{6} \\ \frac{2}{6} \\ \frac{1}{6} \end{pmatrix} \equiv \begin{pmatrix} 0.1667 \\ 0.3333 \\ 0.3333 \\ 0.1667 \end{pmatrix}. \tag{4.4}$$

We show that I is not obtained by the power method. In fact, it is easy to see

$$I_{2n} = H^{2n}\mathbf{e}_1 = \begin{pmatrix} a_n \\ 0 \\ c_n \\ 0 \end{pmatrix} \text{ and } I_{2n+1} = H^{2n+1}\mathbf{e}_1 = \begin{pmatrix} 0 \\ b_n \\ 0 \\ d_n \end{pmatrix}, \tag{4.5}$$

where we do not need exact values of a_n, b_n, c_n, d_n. As a result, $\{I_n\}$ does not converge.

Next, as a trial we change the initial vector. For example we set

$$J_n = H^n J_0 \text{ and } J_0 = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} \tag{4.6}$$

because of $H^T J_0 = J_0$. Let us list the calculation:

$$J_{10} = \begin{pmatrix} 0.1667 \\ 0.3333 \\ 0.3333 \\ 0.1667 \end{pmatrix}, \; J_{11} = \begin{pmatrix} 0.1666 \\ 0.3334 \\ 0.3334 \\ 0.1666 \end{pmatrix}, \; J_{12} = \begin{pmatrix} 0.1667 \\ 0.3333 \\ 0.3333 \\ 0.1667 \end{pmatrix}. \tag{4.7}$$

From the result $n = 10$ is enough.

Last, we present an important.

Problem. We speculate that $J_0 = (1/n, 1/n, \; ..., 1/n)^T$ is in general better than $\mathbf{e}_1 = (1, 0, ..., 0)^T$ as an initial vector. Study this point in detail.

Acknowledgements

We would like to thank Yasushi Homma and Ryu Sasaki for useful suggestions and comments.

References

[1] David Austin, How Google Finds Your Needle in the Web's Haystack, Feature Column, Monthly Essays on Mathematical Topics.
http://www.ams.org/samplings/featurecolumn/fcarc-pagerank.

[2] Semyon Gerschgorin, Über die Abgrenzung der Eigenwerte einer Matrix, Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk 6 (1931), 749-754.

[3] Ichiro Satake, Linear Algebra, Shokabo, Tokyo, 1975 (in Japanese).

[4] Kazuyuki Fujii, Introduction to linear algebra, Lecture Note at Yokohama City University, 2014 (in Japanese).