Application of Matrices in Google Search Engine

Minakshi Panchal[1], Vedant Nemade[2], Nisarga Kasar[3], Pranjal Jadhav[4], Vaishnavi Jadhav[5]
[1, 2, 3, 4, 5] Pimpri Chinchwad College of Engineering, Nigdi, Pune, Maharashtra

Abstract- *Google search engine plays a vital role in our day-to-day life.In the following paper, the answers to - how exactly Google search engine displays relevant web pages, is being discussed which is solely based on the concept of PageRank. Concept of Page rank is the 'heart' of Google search engine .This is one of the application of linear algebra, matrices and their eigenvalues.*

Using some examples this algorithm is explained in detail.

Keywords- PageRank, eigenvalues, eigenvectors, Google search engine.

I. INTRODUCTION

In today's tech-savvy era, Google search engine makes our life much easier than it used to be. Surfing through Google engine does wonders in many ways. Although we just use it on the tip of our fingers many people are not aware about the mechanism that Google search engine follows while answering to our queries on the internet.

Any type of search engine is a system of software, this kind of system is designed to perform web searches that means to search the data on a wide level in a systematic way,and moreover it answers to our textual queries too.

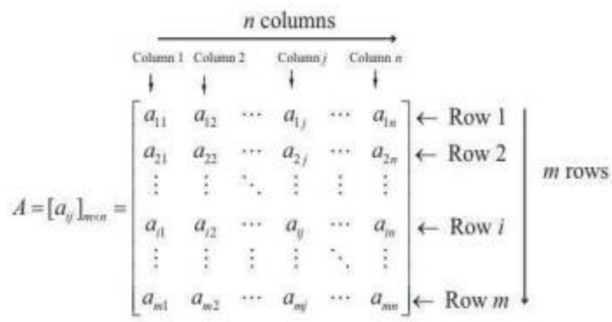
For example: A Google search engine is one of the search engines that is provided by Google and handles over 3.5 billion of searches each day .It comprises around 92% of share for the global search engine market. It is a widely used and abundantly visited website all around the globe.

PageRank is a system of ranks, the order of search results returned by Google is the main theory behind PageRank. In daily life it is visited by a number of people who are either learning or innovating some thing or the other.

So in this paper we are going to study the concept of matrices for its application in search engines.

What is a matrix?

Matrix is a systematic and proper arrangement of any numbers or symbols in the rows and column format. An m×n matrix has m horizontal rows and n vertical columns. The concept of matrices was in discussions from 10th-2nd century BCE from a renowned Chinese text titled ‘The Nine Chapters on the Mathematical Art'. The term matrix was coined by the famous James Joseph Sylvester in 1850.



APPLICATIONS OF MATRICES

Matrices have wide applications in our technical life, branches of engineering, physics, economics, and statistics as well as in various branches of mathematics. Some of them are graph theory (network theory here we have a matrix just having two values in it 1 and 0 and this application can be generally applied to the hyperlinks), cryptography(encryption of data so that only a particular person can see or visit the data and relate information),wireless communication, science, geography, mathematics, probability theory and statistics.

II. MATHEMATICAL BACKGROUND

What is a matrix?

Matrix is a systematic arrangement as mentioned earlier or array of numbers or symbols in the rows and column format. An m×n matrices have m horizontal rows and n vertical columns.

HISTORY OF EIGENVECTORS

The Eigen concept arose from the study of quadratic form and differential equations. Herman von Helmholtz in 1904 used the word eigenvalues and eigenvectors discovered this concept .In 1929 numerical algorithm was published by Richard von Mises for computing eigenvalue and eigenvectors.

The Eigen vector also known as the 'Characteristic Vector' is a non-zero vector which changes by some scalar transformation. Eigen vector to a corresponding non-zero eigenvalue points in a particular stretched direction.

How to calculate the eigenvalues and their eigenvectors?

If A is any square matrix then a non-zero vector v is an eigenvector of A if $Av = kv$ where for some number k called the corresponding eigenvalue.

STEP 1: Set the characteristic equation using $|A - kI| = 0$

STEP 2: Solve the characteristic equation which gives two eigenvalues for 2×2 systems and three eigenvalues for 3×3 systems.

STEP 3: Substitute the respective values into the equation $[A - kI]X = 0$

STEP 4: We obtain suitable eigenvectors from these equations.

Hence the resultant eigenvectors from the corresponding eigenvalues are obtained.

III. LITERATURE REVIEW

1) F MCSHERRY

Done mainly by foundation of an algorithm known as Pagerank an application of linear algebra and matrix computations. Google uses modification of simple power methods or are more and well advanced methods for computing eigenvectors are out of scope due to size of matrix.

2) JONATHAN MACHADO RESEARCH PAPER:

THE paper is based on eigenvalues and eigenvectors and it's application to data problems mainly in google search engine. Eigen vectors are particular vectors that are not rotated by any transformation matrix, and eigenvalues are the amount by which the eigenvectors are stretched further. It tells us about slight relation between these algorithms with Markov chains

.3) K AVRACHENKOV,N LITVAK

Nowadays, Google is the most widely used search engine. Earlier search enginesused simple keyword techniques due to which a lot of irrelevant data used to appear to overcome this problem the founder of google Larry Page and Sergey Brin came up with the idea of PageRank algorithm.

4) AMY N. LANVILLE and CARL D. MEYER

In this paper the author tells about PageRank and HITS algorithm with taking an interesting examples. A user has to sort through the list of websites in order to find the relevant website.

5) PUNIT PATEL AND KANU PATEL

In this paper the authors have compared two popular search engine algorithms namely PageRank and HITS algorithm. It talks about Web content mining, explains how it is responsible for exploring proper information and important data from the contents of the webpages. World Wide Web can be used to extract traffic patterns which is known as Web usage mining.

6) HERBERT S. WILF

In this paper the author tells about the importance of the website and how to calculate it .The author mentions that the importance of a website A is proportional to the addition of importance of all the websites which are pointing towards website A.

7) KURT BRYAN† AND TANYA LEISE‡

This was an interesting study for the analysis of various algorithms or methodologies that reflect on how exactly a web locates it's relevant web pages. Most of the magic in google search engine is done by the important scores and the ranking system. The web ranking problem is converted to standard problem of solving an eigenvectors for the column stochastic matrix

Use of PageRank for Google search Engine

In our everyday life we face a lot of circumstances where we have to decide an appropriate and relevant path. Similarly, when we work with Google search engine the topic or query that we search displays only the appropriate and relevant website links. This is done by the search engine itself in no time. In simple words the algorithm of calculating the number of related and relevant websites and webpages and displaying them in the ascending order of their preferences is known as the PageRank algorithm.

PageRank is the system used for ranking of web pages that Google's well-known founders named Larry Page & Sergey Brin developed at Stanford University in California.

For ranking of the web pages and website we assemble their preferences in the form of a matrix. This matrix is unique for each and every search text that we search.

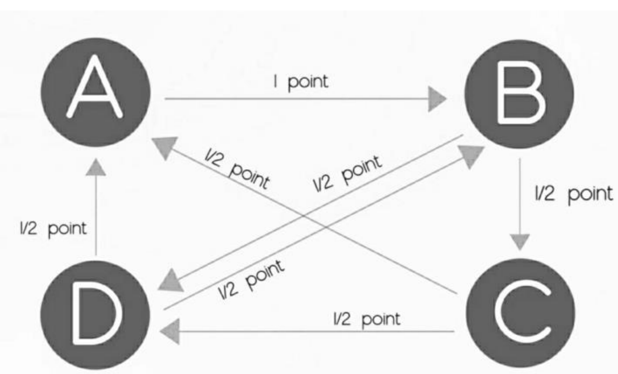
Google has millions of websites and links stored in its memory. When we search something, Google has to display us the required and expected websites, but as stated initially there are no. of websites, which one to be displayed when. In simple language these websites that are related to what we search on Google form a linking relationship with one another and each website has freedom to link to any other website by distributing it's probability of being displayed amongst the other websites to which it links to. While linking the path these websites obeys some basic rules.

One of these rules is that no single website can rank itself but it can rank more than one website simultaneously.

Let us study the concept of page ranking algorithm and the application of eigenvectors and eigenvalues used by the search engine.

Case- 1

Let us assume that a webpage has only 4 websites named as A,B,C& D, these websites now start to give importance and ranking to each other except for itself. Each website has a limit of 1 point that will be distributed or passed on to the other websites via links.



The matrix say X for the above linkage relation can be given as

$$X=\begin{array}{c} \\ \end{array}\begin{matrix} A & B & C & D \end{matrix}$$

$$X=\begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Now, here the matrix is arranged in such a way that the first row and column corresponds to linking points with A itself .The rows and columns basically give us value for again A,B,C,D respectively.

If the webpage has displayed only 4 websites, it means that the selection of any one of these websites is 0.5 each.

Let Y be the matrix of probability i.e. 0.25 for each website to be picked at first.

$$Y=\begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

Therefore,

$$XY=\begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}\begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}=\begin{bmatrix} 0.25 \\ 0.375 \\ 0.125 \\ 0.25 \end{bmatrix}$$

This obtained matrix is again multiplied with X,

$$X(XY)=\begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}\begin{bmatrix} 0.25 \\ 0.375 \\ 0.125 \\ 0.25 \end{bmatrix}$$

The obtained matrix is then again multiplied by the matrix X up to the point of convergence of the matrix.

$$X^{n}Y=\begin{bmatrix} 0.217 \\ 0.348 \\ 0.174 \\ 0.261 \end{bmatrix}$$ where n tends to infinity.

Here, Dhas a link incoming from B and B has the highest value of the Eigen vector therefore although A and D have almost equal eigenvector values, D overshadows A because it is preferred by B .

Hence B is ranked highest.

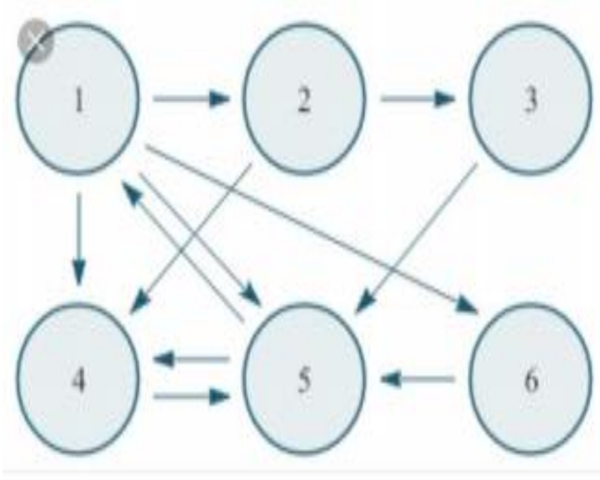
This is of the form $X(X(X \dots \dots (XY)))$.

Adding websites and links to the web just leads to change of numbers, values and dimensions of the matrix.

Let us consider another example and study it in detail.

Case- 2

Suppose the webpage initially displays 6 websites.
The linking chart is as follows:



Let us check for no 1 page which has 4 outgoings(links to pages 2,4,5 and 6).Check that in first column of "links matrix" ,1/4 value is placed in each row 2,4,5,6 since each link is 1/4 th of all outgoing links. Remaining rows in column 1 have value 0,since no 1 page doesn't link to them. Also page 2 has 2 outgoing links to pages 3 and 4. In column 2 we place 1/2 in rows 3 and 4,and 0 in the rest. Same process is carried out for 6 pages.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 0 & 0 & 1/2 & 0 \\ 1/4 & 0 & 1 & 1 & 0 & 1 \\ 1/4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$|A-\lambda I| = \begin{bmatrix} -\lambda & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/4 & -\lambda & 0 & 0 & 0 & 0 \\ 0 & 1/2 & -\lambda & 0 & 0 & 0 \\ 1/4 & 1/2 & 0 & -\lambda & 1/2 & 0 \\ 1/4 & 0 & 1 & 1 & -\lambda & 1 \\ 1/4 & 0 & 0 & 0 & 0 & -\lambda \end{bmatrix}$$

$$= \lambda^6 - \frac{5}{8}\lambda^4 - \frac{\lambda^3}{4} - \frac{\lambda^2}{8}$$

On further solving we will get the eigenvalues,

λ =-0.72031, -0.13985+-0.39240j, 0, 1

but as we are interested in ranking the matrix in order to evaluate page rank ,only the non-negative real values of λ,hence we conclude $\lambda = 1$.

Therefore the eigenvector corresponding to eigenvalue=1 can be calculated using computer algebra system and is as follows:

$$V = \begin{bmatrix} 4 \\ 1 \\ 0.5 \\ 5.5 \\ 8 \\ 1 \end{bmatrix}$$

These values of elements in the matrix v represent the ranking or we can say preferable rank for that corresponding page,8 is the highest value rank amongst all the other webpages. As a result, the 5th page has the most relevant results to our search text and will be displayed second to none, that is it is ranked at the top of searched results. This is how Google search engine works. It works with more optimized mathematical and logical techniques because Google search engine has to deal with billions of web pages and their corresponding links at the same time.

Currently, Google uses over 200 different webpage signals for analyzing the ranking factor including the speed of the page whether local or not, mobile friendliness ,relevance, authority of overall website ,amount of text, number of websites, freshness of the content & so on. Moreover it constantly keeps revising this algorithm and updating it to its better version than the previous one.

IV. CONCLUSION

Matrices is one of the important mathematical tool which is widely used in different applications. In current paper use of Eigen values and Eigen vector is explained for to decide the PageRank. Further Use of The PageRank algorithm works by giving individual web pages a rank, determined by the number of links that are pointed towards the page. The more websites that link to a site, the more valuable the content of that site is considered, and the higher its rank.

REFERENCES

[1] Jonathan Machado, Linear Algebra Application: Google Pagerank Algorithm, Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC 27402, United states of America
E-mail address: jomachad@uncg.edu
[2] Herbert S. Wilf,, Searching the web with eigenvectors
, Philadelphia, PA 19104-6395 April 13, 2001
[3] K. Avrachenkov, N. Litvak, The effect of new links on Google PageRank, Stochastic Models, Volume 22, 2006 - Issue 2, 319-331
[4] Punit Patel And Kanu Patel, A Review of PageRank and HITS Algorithms, International Journal of Advance Research in Engineering, Science & Technology(IJAREST), ISSN(O):2393-9877, ISSN(P): 2394-2444, Volume 2,Issue 1
[5] A. Langville, C. Meyer, A survey of eigenvector methods for Web information retrieval, SIAM Review 47, 1 (2005), 135-161.
[6] Kurt Bryan And Tanya Leise‡ The $25,000,000,000∗ Eigenvector The Linear Algebra Behind Google
www.rose-hulman.edu/∼bryan
[7] F. McSherry, A uniform approach to the accelerated PageRank computation, Proceedings of the 14th International Conference on WWW (2005), 575-582