THE THEORY BEHIND GOOGLE SEARCH ENGINE

INSTITUTE TECHNICAL SUMMER PROJECT
IIT BOMBAY

By

PULKIT SAREEN

MENTOR

SANAT ANAND

Contents

ABSTRACT :
https://docs.google.com/document/d/1CvSxJ5lJ4516Ckui0wURtDAeTspsK_X-VYSBm-IS5LE/edit?usp=sharing

INTRODUCTION

When Google went online in the late 1990's, one thing that set it apart from other search engines was that it's search result listings would always deliver the "valuable stuff" first. With other search engines, you had to browse through various pages to get the appropriate data, because they used to display the links merely by the matched text. Part of the magic behind this is the Page Rank Algorithm used by Google, which quantitatively rates the importance of each page on the web, allowing Google to rank the pages and thereby present the user the more important, most relevant and helpful page first.

Understanding Page Rank algorithm is essential for anyone designing a webpage that they want people to access frequently, since getting listed first in a Google search leads to many people looking at your page. Definitely, due to Google's prominence as a search engine, its ranking has a deep influence on the development and structure of the internet. It also largely influences the kind of information and services that gets influences most frequently.

My idea here idea here is to explain the basic and core idea behind how Google calculates web page rankings. It turns out to be a delightful application of linear algebra.

Google Search uses three basic steps:

1) Crawl the Web using Spiders and locate all web Pages with public access.
2) Index the Pages from the 1st step, so that it can be searched efficiently for relevant keywords and phrases.
3) Rate the importance of each page in the database, so that when a user does a search and the subset of pages in the database with the desired information has been found, the more important pages will be displayed first.

My project basically focuses only on Step 3 and parts of Step 1.

The rated importance of web pages using the page rank algorithm is not the only factor on which the links are displayed, but it is one of the most significant ones. There are many other algorithms used by other search engines like Bing or Yahoo.! But Page Rank is the algorithm used by the ------- Billion Dollar company..!

DEVELOPING A FORMULA TO PAGE RANK

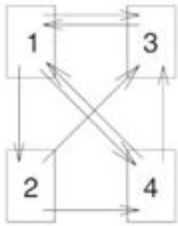

FIG. 2.1. *An example of a web with only four pages. An arrow from page A to page B indicates a link from page A to page B.*

2.1. The basic idea: In what follows we will use the phrase "importance score" or just "score" for any quantitative rating of a web page's importance. The importance score for any web page will always be a non-negative real number. A core idea in assigning a score to any given web page is that the page's score is derived from the links made to that page from other web pages. The links to a given page are called the backlinks for that page. The web thus becomes a democracy where pages vote for the importance of other pages by linking to them.

Suppose the web of interest contains n pages, each page indexed by an integer k, $1 \le k \le n$. A typical example is illustrated in Figure 2.1, in which an arrow from page A to page B indicates a link from page A to page B. Such a web is an example of a directed graph. I'll use x_k to denote the importance score of page k in the web. The x_k is non-negative and $x_j > x_k$ indicates that page j is more important than page k (so $x_j = 0$ indicates page j has the least possible importance score).

A very simple approach is to take x_k as the number of backlinks i.e incoming links for page k. In the example in Figure 2.1, we have $x_1 = 2$, $x_2 = 1$, $x_3 = 3$, and $x_4 = 2$, so that page 3 would be the most important, pages 1 and 4 tie for second, and page 2 is least important. A link to page k becomes a vote for page k's importance.

This approach ignores an important feature one would expect a ranking algorithm to have, namely, that a link to page k from an important page should boost page k's importance score more than a link from an unimportant page. For example, a link to your homepage directly from Facebook ought to boost your page's score much more than a link from, say, bookadda.com. In the web of Figure 2.1, pages 1 and 4 both have two backlinks: each links to the other, but page 1's second backlink is from the seemingly important page 3, while page 4's second backlink is

from the relatively unimportant page 1. As such, perhaps we should rate page 1's importance higher than that of page 4.

As a first attempt at incorporating this idea let's compute the score of page j as the sum of the scores of all pages linking to page j. For example, consider the web of Figure 2.1. The score of page1 would be determined by the relation x1 = x3 + x4. Since x3 and x4 will depend on x1 this scheme seems strangely self-referential, but it is the approach we will use, with one more modification. Just as in elections, we don't want a single individual to gain influence merely by casting multiple votes. In the same vein, we seek a scheme in which a web page doesn't gain extra influence simply by linking to lots of other pages. If page j contains nj links, one of which links to page k, then we will boost page k's score by xj /nj, rather than by xj. In this scheme each web page gets a total of one vote, weighted by that web page's score, that is evenly divided up among its entire outgoing links. To quantify this for a web of n pages, let Lk ⊂ {1, 2, . . . , n} denote the set of pages with a link to page k, that is, Lk is the set of page k's backlinks. For each k we require

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \qquad (2.1)$$

where nj is the number of outgoing links from page j (which must be positive since if j ∈ Lk then page j links to at least page k!). We will assume that a link from a page to itself will not be counted. In this "democracy of the web" you don't get to vote for yourself!

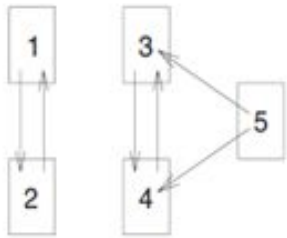

FIG. 2.2. *A web of five pages, consisting of two disconnected "subwebs" W_1 (pages 1 and 2) and W_2 (pages 3, 4, 5).*

Let's apply this approach to the four-page web of Figure 2.1. For page 1 we have x1 = x3 /1 + x4 /2, since pages 3 and 4 are backlinks for page 1 and page 3 contains only one link, while page 4 contains two links (splitting its vote in half). Similarly,

x2 = x1 /3, x3 = x1 /3 + x2 /2 + x4 /2, and x4 = x1 /3 + x2 /2. These linear equations can be written Ax = x, where x = [x1 x2 x3 x4]T and

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}. \qquad (2.2)$$

This transforms the web ranking problem into the "standard" problem of finding an eigenvector for a square matrix! (It is known that eigenvalues λ and eigenvectors x of a matrix A satisfy the equation Ax = λx, x ≠ 0 by definition.) We thus seek an eigenvector x with eigenvalue 1 for the matrix A. We will refer to A as the "link matrix" for the given web.

It turns out that the link matrix A in equation (2.2) does indeed have eigenvectors with Eigenvalue 1, namely, all multiples of the vector [12 4 9 6]T (it is known that any non-zero multiple of an eigenvector is again an eigenvector). Let's call to scale these "importance score eigenvectors" such that the components score to 1. In this case we obtain

x_1=12/31=0.387,

x_2=4/31=0.129,

x_3=9/31=0.290,

x_4=6/31=0.194 .

Point to be noted here is that this ranking differs from the one obtained by simply counting backlinks. It might seem surprising that page 3, linked to by all other pages, is not the most important. To understand this, note that page 3 links only to page 1 and so casts its entire vote for page 1. This, with the vote of page 2, results in page 1 getting the highest importance score.

More generally, the matrix A for any web must have 1 as an eigenvalue if the web in question has no dangling nodes (pages with no outgoing links). To see this, first note that for a general web of n pages formula (2.1) gives rise to a matrix A with Aij = 1/nj if page j links to page i, Aij = 0 otherwise. The jth column of A then contains nj non-zero entries, each equal to 1/nj , and the column thus sums to 1.

Thus the following definition is used to study the Markov Chain.

Definition 2.1: A square matrix is called a **column-stochastic matrix** if all of its entries are non-negative and the entries in each column sum to 1.

Proposition 1: Every column-stochastic matrix has 1 as an eigenvalue.

Proof: Let A be an n x n column-stochastic matrix and e denote an n dimensional column vector with all entries equal to 1. We know that A and $\mathbf{A}^T$ have the same eigenvalues. Since A is column stochastic it is easy to see that $\mathbf{A}^T e = e$, so that the eigenvalue of A and $\mathbf{A}^T$ is 1.

In the following text, V_1 (A) to denote the Eigenspace for eigenvalue 1 of a column-stochastic matrix A.

2.2. Shortcomings. Several difficulties arise with using formula (2.1) to rank websites. In this section we discuss two issues: webs with non-unique rankings and webs with dangling nodes.

3. Non-Unique Rankings: For our rankings it is desirable that the dimension of V1(A) equal one, so that there is a unique eigenvector x with $\Sigma x_i = 1$ that we can use for importance scores. This is true in the web of Figure 2.1 and more generally is always true for the special case of a strongly connected web (that is, you can get from any page to any other page in a finite number of steps.
Unfortunately, it is not always true that the link matrix A will yield a unique ranking for all webs. Consider the web in Figure 2.2, for which the link matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We find here that V1(A) is two-dimensional; one possible pair of basis vectors is x = [1/2, 1/2, 0, 0, 0]T and y = [0, 0, 1/2, 1/2, 0]T . But note that any linear combination of these two vectors yields another vector in V1(A),
eg : 3x/4 + y/4 = [3/8, 3/8, 1/8, 1/8, 0] . It is not clear which, if any, of these three Eigenvectors we should use for the rankings!!

It is no coincidence that for the web of Figure 2.2 we find that the number of solutions i.e dim(V1(A)) > 1. It is a consequence of the fact that if a web W, considered as an undirected graph(where edges are bidirectional, ignoring which

direction each arrows points), consists of r disconnected subwebs W1, . . . ,Wr, then dim(V1(A)) ≥ r, and hence there is no unique importance score vector x ∈V1(A) with Σxi = 1. This makes intuitive sense, if a web W consists of r disconnected subwebs W1, . . . ,Wr then one would expect difficulty in finding a common reference frame for comparing the scores of pages in one subweb with those in another subweb.

Indeed, it is not hard to see why a web W consisting of r disconnected subwebs forces dim(V1(A)) ≥ r. Suppose a web W has n pages and r component subwebs W1, . . . ,Wr. Let ni denote the number of pages in Wi. Index the pages in W1 with indices 1 through n1, the pages in W2 with indices n1 + 1 through n1 + n2, the pages in W3 with n1 + n2 + 1 through n1 + n2 + n3, etc. In general, let Ni = Σ nj from j=1 to i, for all i ≥ 1. with N0 = 0, so Wi contains pages Ni−1 + 1 through Ni. For example, in the web of Figure 2 we can take N1 = 2 and N2 = 5, so W1 contains pages 1 and 2, W2 contains pages 3, 4, and 5. The web in Figure 2.2 is a particular example of the general case, in which the matrix A assumes a block diagonal structure

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_r \end{bmatrix},$$

where Ai denotes the link matrix for Wi . In fact, Wi can be considered as a web in its own right. Each ni × ni matrix Ai is column-stochastic, and hence possesses some eigenvector vi ∈ lRni with Eigenvalue 1.

4. Dangling Nodes: Another difficulty may arise when using the matrix A to generate rankings. A web with dangling nodes produces a matrix A which contains one or more columns of all zeros. In this case A is column-substochastic, that is, the column sums of A are all less than or equal to one. Such a matrix must have all Eigenvalues less than or equal to 1 in magnitude, but 1 need not actually be an eigenvalue for A. Nevertheless, the pages in a web with dangling nodes can still be ranked use a similar technique. The corresponding sub stochastic matrix must have a positive eigenvalue λ ≤ 1 and a corresponding eigenvector x with non-negative entries that can be used to rank the web pages.

5. A remedy for dim(V1 (A)) > 1 : An enormous amount of computing resources are needed to determine an eigenvector for the link

matrix corresponding to a web containing billions of pages. It is thus important to know that our algorithm will yield a unique set of sensible web rankings. The analysis above shows that our first attempt to rank web pages leads to difficulties if the web isn't connected. And the worldwide web, treated as an undirected graph, contains many disjoint components.

6. A modification to the link matrix A: For a n page web or a web consisting of server farms, with dangling nodes we can generate unambiguous importance scores as follows, including cases of web with multiple subwebs.

Let S denote an $n \times n$ matrix with all entries $1/n$. The matrix S is column-stochastic, and it is easy to check that V1 (S) is one-dimensional. We will replace the matrix A with the matrix

$$M = (1 - m)A + mS,$$

where $0 \leq m \leq 1$. M is a weighted average of A and S. The value of m originally used by Google is reportedly 0.15 . For any $m \in [0, 1]$ the matrix M is column-stochastic and V1 (M) is always one-dimensional if $m \in (0, 1]$. Thus M can be used to compute unambiguous importance scores. In the case when m = 0 we have the original problem, for then M = A. At the other extreme is m = 1, yielding M = S. This is the ultimately egalitarian (all equal) case: the only normalized eigenvector x with eigenvalue 1 has $x_i = 1/n$ for all i and all web pages are rated equally important.

Using M in place of A gives a web page with no backlinks (a dangling node) the importance score of m/n , and the matrix M is substochastic for any $m < 1$ since the matrix A is substochastic. Therefore the modified formula yields nonzero importance scores for dangling links (if $m > 0$) but does not resolve the issue of dangling nodes. In the remainder of this article, we only consider webs with no dangling nodes.

The equation $x = Mx$ can also be cast as

$$x = (1 - m)Ax + ms,$$

where s is a column vector with all entries $1/n$. Note that $Sx = s$ if $\Sigma x_i = 1$.

PAGERANK IN SHORT

Page Rank is calculated on the basis of the number of incoming links to the page. It is similar to a vote give by one page to another. But each vote carries a different weight. This weight is determined by the number of pages it links to.

Several difficulties such as Non-Unique ranking arise in the Web where all web pages are not connected.

For non-connected web pages, a new diagonal link matrix is created, assuming all the non-connected web pages as separate links.

For dangling nodes (pages with no outgoing links), a new matrix is defined such that it is the weighted average of the link matrix and the matrix with entries 1/n , n is the total number of web pages and the order of the matrices is the same. The weights can be set accordingly. Google uses the weights as 0.85 and 0.15 respectively.

7. CONCLUSION

Page Rank is one of the most powerful and accurate algorithms used in search engines to rank data. It can also be used to rank any data that has a graphical structure similar to the web.

Certain tips to have a better Page Rank

Google loves **unique content**. This fact also takes into account the **freshness of a page**. A page should be constantly updated with time.

"Backlinks" are incredibly important because they are basically like "votes" for your page that tell search engines that other webpages like and utilise your page for information. The more websites that you have linked or "voting" for your web page, the higher your ranking will be. Though, not all votes are weighted the same, as links from the homepage of major websites, like the amazon.in or facebook.com, are going to be worth a lot more than a lesser known website. **"Weighted Backlinks"** play a very major role in the ranking of web page.

Your page must also contain **"keywords"** that you expect the user to type in to search your page.

Quality content needs to be provided. If the quality of content provided is good, other pages will want to link to your page, which increases the Pagerank. **Regularly posted** quality content is the recommended method to getting more traffic to your website. Ideally you should post at least 2 new articles per week on a blog, more being better, as Google will favour websites with **frequently updated material** that contains **unique** and **relevant content**.

.

8. BIBLIOGRAPHY

1) http://nptel.ac.in/courses/111106050/
2) Topics In Structural Graph Theory. By : Lowell W. Beineke and Robin J. Wilson (Cambridge Publishers)
3) https://en.wikipedia.org/wiki/PageRank
4) https://www.briggsby.com/methods-for-evaluating-freshness/
5) https://en.wikipedia.org/wiki/Markov_chain