On the Google Pagerank Algorithm

Ryan Joo Rui An

2024

Abstract

PageRank is an algorithm used by Google Search to rank web pages in their search engine results. It is named after both the term "web page" and co-founder Larry Page. PageRank is a way of measuring the importance of website pages.

Contents

1 Introduction

A *search engine* aims to rank web pages effectively and efficiently, ensuring that the most relevant sites appear at the top of the search results.

To formulate such a ranking algorithm, we make the following assumptions:

- More important (authoritative) sites are likely to receive more links from other sites.
- The probability of transitioning from one site to another is uniform; that is, a person randomly refers to other sites with equal probability.

2 Mathematical Model

2.1 Graph Theory

Let S be a set containing sites s_i $(1 \leqslant i \leqslant n)$ which that contain a certain keyword. That is,

$$S = \{s_1, s_2, \ldots, s_n\}.$$

When a site s_i references another site s_j, we can draw a directed edge from node i to node j. This gives a *directed graph* $G = (V, E)$, where V is the set of nodes and E is the set of directed edges;

$$E = \{(i, j) \mid s_j \text{ references } s_i\}.$$

The directed graph G is known as the *web graph*; see Fig. 1.

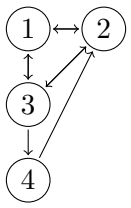

Figure 1: A directed graph G representing the web graph.

We can represent the web graph as an adjacency matrix.

Definition 2.1 (Adjacency matrix)**.** An *adjacency matrix* is a matrix $A = (a_{ij})_{n\times n}$ whose entries are defined as

$$a_{ij} = \begin{cases} 1 & (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Then the adjacency matrix A of the web graph in Fig. 1 is given by

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

2.2 Markov Chains

Definition 2.2 (Probability vector)**.** A *probability vector* is a vector with non-negative entries that add up to 1.

A *Markov chain* is a mathematical model that describes an experiment or measurement that is performed many times in the same way, where the outcome of a given experiment can affect the outcome of the next experiment. The process starts at an initial state described by a probability vector $\mathbf{x}_0$, and transitions successively from one state to another, say $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. The outcome of a given state depends only on the immediately preceding state. [1]

Definition 2.3 (Column-stochastic matrix)**.** A *column-stochastic matrix* is a square matrix in which all entries are non-negative, and whose columns are probability vectors.

The column-stochastic matrix P is given by

$$p_{ij} = \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}},$$

where we divide each entry in the adjacency matrix A by the sum of the entries in the corresponding column, such that the columns of the matrix are now probability vectors.

Then the column-stochastic matrix P of the web graph in Fig. 1 is given by

$$P = \begin{pmatrix} 0 & 1/2 & 1/3 & 0 \\ 1/2 & 0 & 1/3 & 1 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \end{pmatrix}.$$

We now want to analyse the long-term behaviour of the Markov chain after starting at some initial state. Thus, a Markov chain can be expressed as the following recurrence relation:

$$\mathbf{x}_{k+1} = P\mathbf{x}_k \quad \text{for } k = 0, 1, 2, \ldots \tag{1}$$

To compute $\mathbf{x}_k$ in general, repeated application of (1) gives

$$\mathbf{x}_k = P^k \mathbf{x}_0. \tag{2}$$

In the long run, it is expected that the state vector will converge to a unique vector, which is independent of the initial state vector $\mathbf{x}_0$. This is known as the *equilibrium vector*, denoted by $\mathbf{x}_\infty$. That is, taking the limit as $k \to \infty$ in Eq. (2) gives

$$P\mathbf{x}_\infty = \mathbf{x}_\infty. \tag{3}$$

2.3 Eigenvectors and Eigenvalues

Let $\mathcal{M}_{m\times n}(\mathbf{R})$ denote the set of all $m \times n$ matrices with real entries. We now introduce the concepts of eigenvectors and eigenvalues.

Definition 2.4 (Eigenvalue and eigenvector)**.** Let $A \in \mathcal{M}_{m\times n}(\mathbf{R})$, and let $\mathbf{v} \in \mathbf{R}^n$, $\mathbf{v} \neq \mathbf{0}$ for which

$$A\mathbf{v} = \lambda\mathbf{v} \tag{4}$$

for some scalar $\lambda \in \mathbf{R}$. Then λ is called an *eigenvalue* of the matrix A, and $\mathbf{v}$ is called an *eigenvector* of A associated with λ.

Thus Eq. (3) implies that the equilibrium vector $\mathbf{x}_\infty$ is the eigenvector of the column-stochastic matrix P associated with eigenvalue 1.

Hence, when given the column-stochastic matrix P, we can compute the equilibrium vector $\mathbf{x}_\infty$ associated with eigenvalue 1. The sites are then ranked according to the entries in the equilibrium vector $\mathbf{x}_\infty$. Thus the most important sites are those with the highest entries in the equilibrium vector.

We now discuss how to compute the equilibrium vector $\mathbf{x}_\infty$. Suppose $\mathbf{x}_\infty$ satisfies Eq. (4). Then

$$P\mathbf{x}_\infty - \mathbf{x}_\infty = 0,$$

or

$$(P - I)\mathbf{x}_\infty = 0,$$

where I is the identity matrix. Equivalently we write

$$(I - P)\mathbf{x}_\infty = 0, \tag{5}$$

which is more commonly used. To solve for $\mathbf{x}_\infty$, we solve the homogeneous system of linear equations in Eq. (5).

Referring to the column-stochastic matrix P in Fig. 1, we have

$$I - P = \begin{pmatrix} 1 & -1/2 & -1/3 & 0 \\ -1/2 & 1 & -1/3 & -1 \\ -1/2 & -1/2 & 1 & 0 \\ 0 & 0 & -1/3 & 1 \end{pmatrix}.$$

Let $\mathbf{x}_\infty = (x_1, x_2, x_3, x_4)$. Then the system of equations in Eq. (5) is

$$\begin{aligned} x_1 - \frac{1}{2}x_2 - \frac{1}{3}x_3 &= 0, \\ -\frac{1}{2}x_1 + x_2 - \frac{1}{3}x_3 - x_4 &= 0, \\ -\frac{1}{2}x_1 - \frac{1}{2}x_2 + x_3 &= 0, \\ -\frac{1}{3}x_3 + x_4 &= 0. \end{aligned}$$

Solving the system of equations, we obtain the equilibrium vector

$$\mathbf{x}_\infty = \begin{pmatrix} \frac{3}{11} \\ \frac{4}{11} \\ \frac{3}{11} \\ \frac{1}{11} \end{pmatrix}.$$

Thus the sites are ranked as follows, in decreasing order of importance:

$$s_2, \quad s_1, \quad s_3, \quad s_4.$$

3 Conclusion

The Pagerank algorithm is a powerful tool that allows search engines to rank web pages effectively and efficiently. By leveraging the linking structure of the web, the algorithm can determine the relative importance of each page, ensuring that the most relevant sites appear at the top of the search results.

References

[1] J. Machado. “Linear Algebra Application: Google PageRank Algorithm”. In: (2019).