UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Matemàtiques

mapthe

November 15, 2022.

PageRank: a different point of view

A. Carmona, A.M. Encinas, M.J. Jimenez, M. Mitjana
Departament de Matemàtiques, UPC.

Abstract

To compute PageRank in the classical model, it is supposed that for some fixed probability d, a surfer jumps to a random node with probability d (damping factor) and goes to an adjacent node with probability $(1-d)$. In the personalized PageRank, a vector v (teleportation or personalized vector) is also considered. Then, the personalized PageRank is the unique probability eigenvector of the Google matrix associated with the eigenvalue 1. The Google matrix, see [3], is

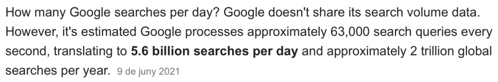
$$G = (1-d)P + d\,\mathsf{e}\,\mathsf{v},$$

where P is the transition probability matrix and e is the all one vector. Some methods to compute the PageRank consider the M-matrix $\mathsf{I}-\mathsf{G}$, which is singular and weakly diagonally dominant. Other models consider also a constant probability of remaining in the node, the so-called lazy parameter that correspond to consider $\dfrac{\mathsf{I}+\mathsf{P}}{2}$ instead of P, then $I-G$ is a diagonally dominant M-matrix and hence it is nonsingular.

G is stochastic it is convex combination of the two stochastic matrices S and $\mathsf{e}\mathsf{v}^\top$ The fundamental centrality measure PageRank implicitly uses Schrödinger operators for its formulation, which corresponds to use diagonally dominant M-matrices. This is due to the presence of the damping parameter for the formulation of the ranking process. Therefore, it is possible, to extend this centrality measure to general Schrödinger operators; that is, to general M-matrices. We plan here to tackle a more realistic model with a wider range of applications. Specifically, we consider in each step of the random walk the importance of both the present state and the state we want to reach. Moreover, the lazy term can be considered as a function instead of a parameter. This model appears when considering a transition probability matrix associated with a symmetric M-matrix (singular or not singular); that is, we can erase the diagonally dominant hypothesis.

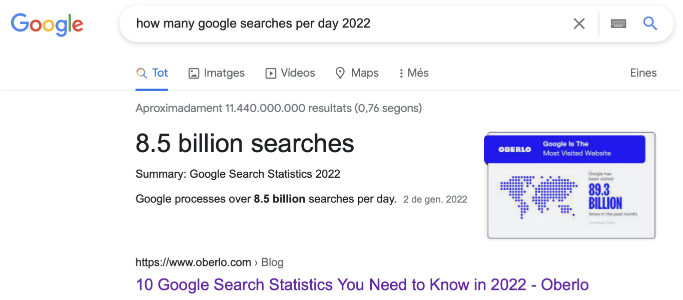
1 Introduction

The rapid growth of the World Wide Web has created a need for search tools.
(Let's ask Google, how many searches does daily)
The first we found is that Google doesn't share its search volume data, (2021)



How many Google searches per day? Google doesn't share its search volume data. However, it's estimated Google processes approximately 63,000 search queries every second, translating to **5.6 billion searches per day** and approximately 2 trillion global searches per year. 9 de juny 2021

...but I succeded to find a more recent data that says the number of Google searches per day is 8.5 billion (american bilion) this is 8.5 times ten to nine.



Aproximadament 11.440.000.000 resultats (0,76 segons)

8.5 billion searches

Summary: Google Search Statistics 2022

Google processes over **8.5 billion** searches per day. 2 de gen. 2022

https://www.oberlo.com › Blog

10 Google Search Statistics You Need to Know in 2022 - Oberlo

This is aproximately almost ten thousand milion of searches per second, or one daily search per inhabitant of the hole world.

$$8.5 \times 10^9 \frac{\text{searches}}{1\text{day}} \frac{1\text{day}}{24\text{hours}} \frac{1\text{hour}}{3600\text{second}} = 98379,63 \frac{\text{searches}}{1\text{second}}$$

2 Google's PageRank algorithm

Why is it used by most of us? Why is it so efficient? One of the best-known algorithms in web search is Google's PageRank algorithm that consists in obtaining a vector raiting the importance of web pages. This vector contains the information of the long term behavior of the random surfer, for instance if the value of the PageRank in pages i and j is respectively 0.35 and 0.01 it means that the web surfer has visited page i 35% of the time and 1% of it for page j. Thus, page i is more important than page j

We first model the web as a directed graphs whose nodes are each web page and there is a link from node i to node j if it is possible to jump from page i to page j.

3 The random surfer model

Imagine a web surfer who bounces along randomly following the hyperlink structure of the web. That is, when he arrives at a page with several outlinks, he chooses one at random, hyperlinks to this new page, and continues this random decision process indefinitely. In the long run, the proportion of time the random surfer spends on a given page is a measure of the relative importance of that page. If he spends a large proportion of his time on a particular page, then he must have, in randomly following the hyperlink structure of the Web, repeatedly found himself returning to that page. Pages that he revisits often must be important, because they must be pointed to by other important pages. Unfortunately, this random surfer encounters **some problems**. He gets caught whenever he enters a dangling node.

Therefore, the Page Rank convergence problems caused by sinks and cycles can be overcome if H is modified slightly so that it is a Markov chain with the desired properties stochasticity adjustment because the $0^\top$ rows of H are replaced with $1/n\mathsf{e}^\top$ As a result, the random surfer, after entering a dangling node, can now hyperlink to any page at random.

Moreover the surfer can get bored or become "lazy" o stop jumping for a while, the meaning of parameter α is: the random surfer keeps traveling through the web with probability α and with probability $1 - \alpha$ decides to remain.

It is known from the theory of Markov chains that for any starting vector the power method applied to a a Markov matrix converges to a unique positive vector called the stationary vector as long as P is stochastic, irreducible and aperiodic.

That is $S = \mathsf{H} + \mathsf{d}\mathsf{v}^\top$ The random surfer argument for the primitivity adjustment goes like this. While it is true that surfers follow the hyperlink structure of the Web, at times they get bored and abandon the hyperlink method of surfing by entering a new destination in the browser's URL line. When this happens, the random surfer, like a Star Trek character, "teleports" to the new page, where he begins hyperlink surfing again, until the next teleportation, and so on. To model this activity mathematically, Brin and Page invented a new matrix G, such that $\mathsf{G} = \alpha\mathsf{S} + (1-\alpha)\frac{1}{n}\mathsf{e}\mathsf{e}^\top$, where α is a scalar between 0 and 1. G is called the Google matrix. In this model, α is a parameter that controls the proportion of time the random surfer follows the hyperlinks as opposed to teleporting. Suppose $\alpha = .6$. Then 60% of the time the random surfer follows the hyperlink structure of the Web and the other 40% of the time he teleports to a random new page.

The nonzero elements of row i correspond to the outlinks of page i whereas the non zero elements of column i correspond to inlinks pages of page i.

Define the PageRank vector, a vector holding the global measure of importance for each page, to be the stationary vector for a Markov chain related to P [2, 3]. This definition is intuitive, as the stationary vector gives the long-run proportion of time the chain will spend in each state.

In the random surfer model let us look at the combinatorial Laplacian that may be expressed as the difference between a diagonal matrix (degrees) and the (generalized) adjacency matrix. so L is irreducible iff the network is connected.

From it we can obtain Δ the probabilistic Laplacian by dividing each row by the degree, so that the Δ is equal to the identity minus a transition probabilistic matrix, an stochastic irreducible an aperiodic(??).

So we know how to associate P with L which is irreducible, symmetric and diagonally dominant, and $\lambda = 0$ is the lowet eigenvalue and $\omega = 1$ is the associated eigenvector.

Getting inspired with this fact, we consider a generalized $\mathsf{P}^{\lambda\omega}$ associated with L_q an irreducible symmetric M-matrix whose lowest eigenvalue is λ and ω the corresponding eigenvector. By the way, according to Perron Frobenius' theory, ω is a positive vector.

Who is L_q? is a positive semidefinite Schordinger matrix that can be written as the sum of a Laplacian of a certain network and a diagonal matrix Observe that there is no need for L_q to be diagonally dominant.

Characterization of M matrices. This is a well know result, I would like just mention.
As we related a transition probability matrix with a Laplacian matrix we now relate an Schrodinger matrix with a transition probability matrix with respect to λ an ω because the effective resistance is related with the escape probability for a reversible Markov chain. Therefore, the effective resistance with respect to a non-negative value and a weight will correspond to a generalization of the escape probability.

4 Questions/answers?

The advantage of using Schrödinger matrices, this is Schodinger operators is that we can raise BVP in the sense that we can consider absorbing states or lazy nodes. This situation may correspond to study not only Poisson problems, but Neuman or Dirichlet problems or mixed.
If we associate the Schoringer matrix to a transition probability matrix w.r.t. λ and ω, which is the role of the parameter λ is it like the responsible of teledeportation.
See [6], [7],[3],[4], [2], [5], [1].

References

[1] E. Bendito, A. Carmona, A.M. Encinas, J. Gesto, and M. Mitjana. Kirchhoff Indexes of a network. *Linear Algebra and Applications*, 432:2278–2292, 2010.

[2] Abraham Berman and Robert J Plemmons. *Nonnegative matrices in the mathematical sciences*, volume 9 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[4] C. S. H. Calva and A. P. Riascos. Optimal exploration of random walks with local bias on networks. *Physical Review E*, 105(4):1–9, 2022.

[5] A.M. Langeville and C D Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*, volume 59. 2011.

[6] S. Vigna. Spectral ranking. *Network Science*, 4(4):433–445, 2016.

[7] Yue Xie, Ting Zhu Huang, Chun Wen, and De An Wu. An improved approach to the pagerank problems. *Journal of Applied Mathematics*, 2013(1), 2013.