Os fascínios do Google pelas veredas da matemática

D. S. Carvalho, G. M. Souto
Universidade Federal do Espírito Santo

PET MATEMÁTICA UFES

Objetivo do projeto

Temos como objetivo relacionar de forma simples os conceitos matemáticos envolvidos nas nossas buscas pelo Google e o algoritmo PageRank utilizado por este site, que classifica o termo pesquisado pelo grau de importância. Além disso, ainda pretende-se utilizar ferramentas de álgebra Linear para estudar o algoritmo e entender como ocorre essa atribuição da importância nas páginas encontradas que possuem respostas relevantes para nossas pesquisas. Assim, mostraremos de forma prática e cativadora a relação entre a matemática e o nosso cotidiano.

Introdução

Google, trocadilho feito com o termo matemático Googol: número representado pelo dígito um seguido de cem dígitos zeros, ou seja,

$$\mathbf{10^{10^{10}}},$$

deu origem ao nome de uma das maiores e mais evidentes empresas da atualidade: a Google. A utilização da expressão traduz a vontade de Larry Page e Sergey Brin, os fundadores do Google, de organizar um volume muito grande de dados na internet.

Tudo começou na década de 90, com um sistema que foi nomeado BackRub:

$$\text{Googleplex} \Rightarrow \text{Googol} \Rightarrow \text{Google} \Rightarrow \text{Google Search Engine}.$$

O Google utiliza programas chamados "aranhas", os rastreadores da Web, para obter as páginas publicas.

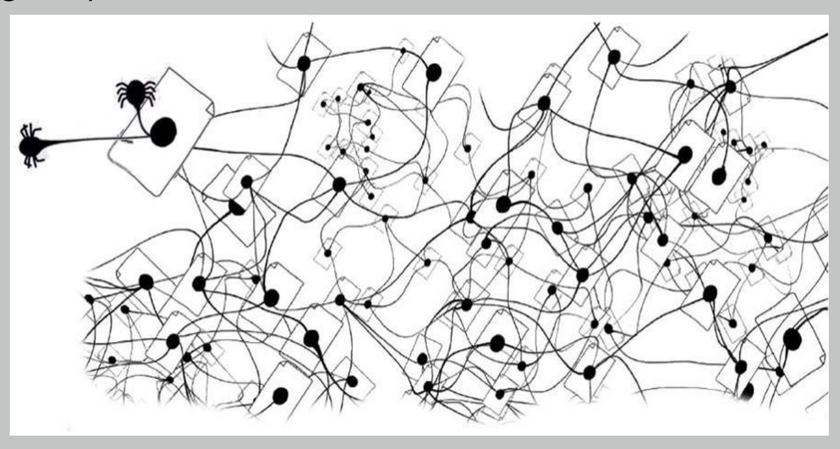

Figure 1: Rastreadores

Distribuição da importância

- Como o Google decide quais resultados mais nos interessam?
  - Com a análise de mais de **200** variáveis, cada uma com um respectivo peso. Algumas dessas variáveis são: Quantidade de vezes que essa página contém suas palavras-chave, posições em que aparecem as palavras pesquisadas na página: no título da página ou no nome do site. Além dessas, o PageRank.
- O que é PageRank? É uma métrica criado pelo Larry Page e utilizada pelo Google dentro do seu algoritmo para entender a importâcia de uma página da Web.
- Pré-requisitos: O que é preciso saber para entender o método?
  - Matrizes: Operações e propriedades;
  - Sistemas lineares: resoluções e classificação através do escalonamento.
  - Teoria dos grafos: um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Para tal são empregadas estruturas chamadas de grafos.

Roteiro

Para a apresentação do algoritmo PageRank, vamos supor que a internet possui quatro páginas A, B, C e D.

- Monte o sistema de equações lineares que indica as relações de importãncia entre as páginas;
- Utilize o conjunto solução do sistema obtido para classificar qual das páginas é a mais importante.

Informações

- Deyze Santos Carvalho, Orientanda.
- Ginnara Mexia Souto, UFES, Orientadora.

Desenvolvimento

- No modelo de internet com **4** páginas, a página C recebe links das páginas A, B e D. Dizemos que a importância de cada página é gerada pela soma das contribuições dadas á ela pelas outras páginas.
- No nosso exemplo com 4 páginas da internet, assim que escalonamos a matriz chegamos a um sistema possível indeterminado, possuindo assim infinitas soluções. Uma das soluções encontradas é: $A = 3, B = 9, C = 16, D = 12.$ Logo, C é a página mais relevante.

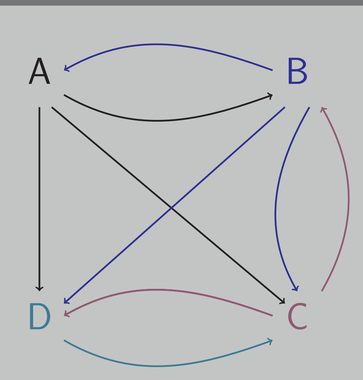

Figure 2: Grafo de links nas paginas.

$$\begin{cases} x_A = \frac{1}{3}x_B \\ x_B = \frac{1}{3}x_A + \frac{1}{2}x_C \\ x_C = \frac{1}{3}x_A + \frac{1}{3}x_B + 1D \\ x_D = \frac{1}{3}x_A + \frac{1}{3}x_B + \frac{1}{2}x_C \end{cases}$$

PageRank para n páginas

No caso de n páginas, teremos um Sistema linear com n variáveis.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = x_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = x_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = x_n \end{cases}$$

Esse sistema pode ser reescrito como $Ax = x$, onde $A = [a_{ij}]_{n\times n}$ e $x = (x_1, x_2, ..., x_n)$. Supomos que cada página possui um link que sai dela para outra página.

No caso do PageRank a matriz A obtida será coluna-estocástica, ou seja, as somas das entradas em cada coluna é igual a $\mathbf{1}$. Com argumentos de Álgebra Linear, provamos que $\mathbf{1}$ é um autovalor de A, donde sempre garantimos a existência de solução para o sistema linear.

Usando o Teorema do Ponto fixo de Banach, podemos obter um método para conseguir a solução x

$$x = \lim_{n\to+\infty} A^n p,$$

para qualquer $p \in \mathbb{R}^n$

Considerações

- No exemplo usando **4** páginas conseguimos resolver por meio do escalonamento ainda de forma prática. Porém, sabemos que existem uma infinidade de links na web, nos dando assim sistemas $n \times n$, onde podemos usar os conceitos de autovalor para descobrirmos essa importância.
  Associando assim vários conceitos que aprendemos em Álgebra Linear para atribuir valores a cada página da internet.

Conclusões

- A internet esta cada vez mais presente no nosso dia a dia, sempre que estamos com alguma dúvida recorremos a ela. É natural querermos entender algo tão presente no nosso cotidiano, desta forma verificar os conceitos matemáticos envolvidos nesta atividade é também apresentar de forma prática como a matemática está presente em tudo. Esperamos que após todos os conceitos introduzidos, o expectador possa reconhecer como algumas das ferramentas da Álgebra Linear podem ser empregadas.

Referências

[1] Autor desconhecido.
ALGORITMOS.
https://www.google.com/intl/pt-BR_ALL/insidesearch/howsearchworks/algorithms.html, 2011.

[2] J. C. BATTI.
Um pouco da matematica por tras do algoritmo pagerank do google.
mestrado profissional em matematica, page 61, 2015.