Introdução à Modelagem Matemática
PageRank

Igor Patrício Michels

22 de novembro de 2021

Boa parte do texto abaixo foi retirado da Wikipedia. Em geral, costumo ver o site em inglês [1], pois as páginas em português costumam ser pouco informativas, como visto em [3]. Entretanto, acho que a melhor referência rápida para PageRank pode ser a página em português da Wikipedia [2].

Introdução

PageRank é um algoritmo utilizado pela ferramenta de busca Google para posicionar websites entre os resultados de suas buscas. O PageRank mede a importância de uma página contabilizando a quantidade e qualidade de links apontando para ela. Não é o único algoritmo utilizado pelo Google para classificar páginas da internet, mas é o primeiro utilizado pela companhia e o mais conhecido.

A ideia do PageRank

Uma ideia básica para classificar resultados de um site é a votação, isto é, aquele mais indicado deve ser o melhor. Mas note que, apesar dessa ideia funcionar, ela pode classificar erroneamente uma boa referência por causa do desconhecimento de alguns sites perante a existência dela.

Exemplo

Perguntando a você o nome de um matemático que você considera relevante você pode prontamente indicar o matemático B. Diversas outras pessoas também acham o matemático B um cara excepcional e o indicam.
Entretanto, ao perguntarmos ao matemático B quem ele considera relevante ele deverá indicar um matemático C, possivelmente alguém que você nunca tenha ouvido falar.
Fazendo a votação até o fim, podemos supor que o matemático C recebeu apenas o voto do matemático B, que por sua vez recebeu vários votos de pessoas que não receberam voto algum ou até mesmo que receberam alguns votos, mas em quantidade inferior a votação de B. A pergunta é: será que o matemático C não é tão relevante quanto o matemático B?
A resposta para essa pergunta é, possivelmente, não. Mas note que o sistema de votação vai induzir a essa conclusão, isto é, o sistema de votação vai induzir a pensarmos que C não é tão relevante quanto B, sendo quase tão importante quanto quem não recebeu votos.

Note que os matemáticos do nosso exemplo podem ser substituídos por qualquer coisa, podem ser jogadores, celebridades, sites, etc. Tradicionalmente a ideia é realizada com sites que linkam a outros sites, então a dúvida passa a ser: será que o site mais linkado realmente é muito mais relevante que o site que ele está linkando?

Figura 1: Será que o site amarelo é mais importante que o site vermelho para o qual ele aponta?

Pensando nisso, a ideia do PageRank é realizar uma "votação", mas com cada voto tendo peso igual a influência do eleitor, isso é, uma pessoa que não recebeu nenhum voto vai ter baixa influência, enquanto quem recebeu muitos votos tem uma influência alta, logo, no nosso exemplo, ao utilizar a ideia do PageRank teremos que o matemático B e o matemático C terão relevância similar.

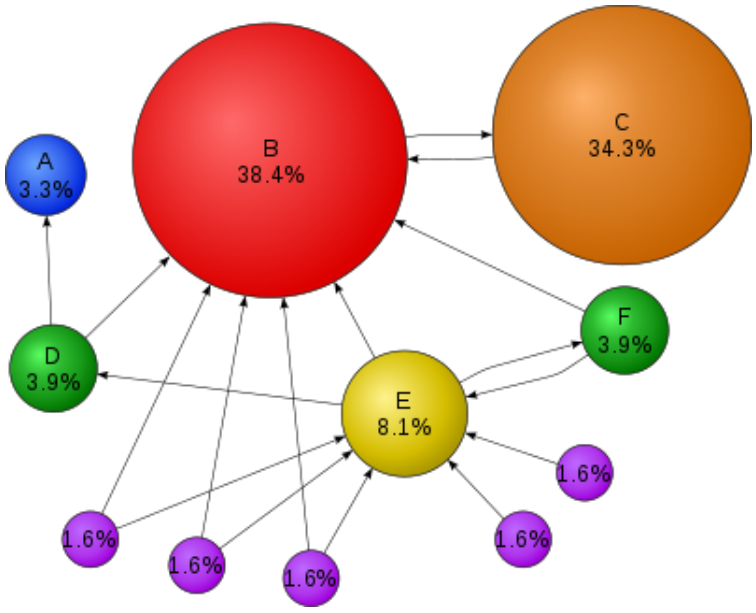

Figura 2: Note que C recebeu apenas um voto, mas esse é um voto de alguém com muita relevância, logo a relevância de B é, em parte, passada para C.

1 O algoritmo

O cálculo do PageRank se dará por meio de um algoritmo iterativo, isso é, realizando o mesmo passo diversas vezes até que haja a convergência.[1]

[1]Por convergência não estamos dizendo, necessariamente, que dois valores consecutivos da sequência serão iguais. Muitas vezes, quando trabalhamos com convergência no âmbito computacional, consideramos que se a diferença entre o resultado de dois passos consecutivos é menor que uma tolerância de, por exemplo 10^{-4}, então houve a convergência.

O passo inicial do algoritmo se dá pela criação de um vetor u_0 atribuindo uma influência igual a $\frac{1}{n}$ para cada nó da rede, onde n representa o número de nós da rede. Nesse caso, a linha i de u_k representa a influência do nó i durante a iteração k. A seguir, poderíamos pensar em ir atualizando esse vetor u_k por meio da equação

$$u_{k+1} = H \cdot u_k,$$

onde H é uma matriz de probabilidades do passeio aleatório dessa rede, isso é, a entrada i, j da matriz H representa a probabilidade de chegar a aresta i quando partimos da aresta j.

Essa ideia é muito boa, mas gera alguns problemas como, por exemplo, nas redes da Figura 3.

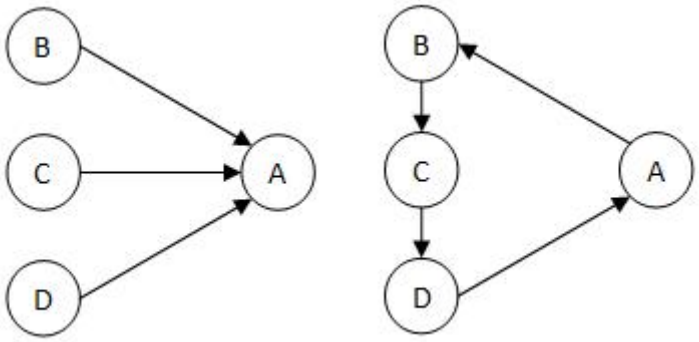

Figura 3: Como funcionaria um passeio aleatório em redes similares a essas?

Note que na rede à esquerda na Figura 3 ficaríamos presos no nó A, enquanto na rede à direita ficaríamos presos num loop, sem visitar outros nós fora desse loop, caso existam. Pensando nesses casos foi inserido um novo parâmetro, d, que representa a probabilidade de se manter no passeio aleatório.[2]

Uma ajudinha na intuição

Com a inclusão desse novo parâmetro d seria equivalente a você ter uma tecla do seu computador que te leva a uma página aleatória da internet. Dessa forma, a ideia é que você esteja navegando na internet e, com probabilidade, $1 - d$ você aperta essa tecla e vai para uma página qualquer (com todas as páginas tendo a mesma probabilidade), enquanto que com probabilidade d você clica em algum link que existe na página atual.

Agora, a modelagem vai mudar um pouquinho, uma vez que, com esse novo parâmetro, podemos ver que a cada iteração a probabilidade de ir para cada página é de, pelo menos, $\frac{1-d}{n}$. Dessa forma, se nossa rede tinha uma matriz de passeio aleatório sendo representada por H, então a nova matriz de passeio aleatório, agora considerando d é

$$\tilde{H} = d \cdot H + (1 - d) \cdot \frac{I}{n},$$

onde temos

- $\tilde{H}$: a nova matriz do passeio aleatório;
- d: a probabilidade de se manter no passeio a cada iteração;
- H: a matriz do passeio original;
- I: a matriz de uns de tamanho $n \times n$ e;
- n: o número de nós da rede.

[2]Esse d também pode ser considerado como um parâmetro de salto, isso é, $1 - d$ representa a probabilidade de darmos um salto na nossa rede.

A matriz $\tilde{H}$ gerada com essa transformação é irredutível, primitiva e estocástica, então é possível provar que a sequência u_k vai convergir, mas não estaremos preocupados com isso nesse curso.

Além disso, é possível interpretar cada um dos termos dessa soma:

- $d \cdot H$ pode ser interpretado como a probabilidade de chegar a um nó por meio do passeio aleatório que já estava sendo descrito e;
- $(1-d) \cdot \dfrac{I}{n}$ pode ser interpretado como a probabilidade de chegar a um nó por meio do "botão" que leva a um nó aleatório da rede.

Feito isso, tendo nossa matriz de transição, $\tilde{H}$, podemos encontrar o PageRank da rede como sendo o vetor para a qual a rede converge.[3] Dessa forma, podemos encontrar o PageRank de uma rede da seguinte forma:

- encontrar a matriz $\tilde{H}$;
- tomar um posição inicial (um vetor u_0 referente a você estar no primeiro nó, por exemplo) e ver para onde a sequência $u_{k+1} = \tilde{H} \cdot u_k$ converge.

Referências

[1] "PageRank". *Wikipedia*. `https://en.wikipedia.org/wiki/PageRank`.

[2] "PageRank - em português". *Wikipedia*. `https://pt.wikipedia.org/wiki/PageRank`.

[3] "PageRank - em italiano". *Wikipedia*. `https://it.wikipedia.org/wiki/PageRank`.

[3] Ou, alternativamente, o autovetor correspondente ao autovalor 1 da matriz, mas isso vocês ainda verão mais adiante no curso de Álgebra Linear.