PageRank e Cadeias de Markov

Cristhian Grundmann
Hanna Rodrigues Ferreira
Igor Cortes Junqueira
Igor Patrício Michels

Dezembro de 2021

Introdução

Imagine que você esteja curioso acerca de um determinado tema, o que você faz? Se a resposta é "dou um Google", parabéns, você acaba de utilizar o algoritmo PageRank! Criado oficialmente em agosto de 1998, a Google surgiu com uma ideia de Larry Page e Sergey Brin. A história se inicia em 1995, quando Page foi conhecer a universidade de Stanford e Brin recebeu a tarefa de mostrar a universidade para ele. No início os dois discordavam sobre muita coisa mas, no ano seguinte, fecharam uma parceria visando criar um mecanismo de busca que pudesse "organizar as informações do mundo e torná-las universalmente acessíveis e úteis". Para isso, desenvolveram um algoritmo que usava os links para determinar a importância de cada página da internet [**?**].

Rapidamente o mecanismo ganhou força, virando o principal mecanismo de buscas. Esse fato se dá por que as estratégias de calcular a relevância, até o momento, eram calculadas usando apenas os dados da própria página, algo que poderia ser facilmente burlado e, com isso, poderia deixar resultados pouco relevantes nas primeiras posições, principalmente com o grande crescimento da internet na época.

Cadeias de Markov

escrever uma historinha aqui

PageRank

Indo um pouco na contramão de algumas outras estratégias, o PageRank visa calcular a relevância de uma página por meio de fatores externos. De maneira simples, podemos usar um grafo para ilustrar uma pequena rede, onde os nós são os sites e as arestas, direcionadas, representam que o site de origem tem um link para o site de destino. É esperado que o site que receba a maior quantidade de links (maior grau de entrada) deve ser o mais relevante, mas e em caso de empate? A ideia de Page e Brin foi ponderar os votos de acordo com a relevância, isso é, se um site tem uma alta relevância, seu voto deve ter um peso maior que o voto de um site que não recebe link algum.

Exemplo

Um site que recebe apenas um link, mas do G1, deve ser mais relevante que um site que recebe apenas um link de um site pessoal.

Conexão com as Cadeias de Markov

Intuitivamente, podemos pensar no processo de um internauta ficar navegando na internet e trocar de sites por meio de links presentes no próprio site, com mesma probabilidade para cada link. Dessa forma, se

um site S tem link para n diferentes sites (T_1, T_2, ... e T_n), a probabilidade do internauta sair do site S para os sites $T_i, i \in \{1, 2, \ldots, n\}$ é igual a $\frac{1}{n}$ e é nula para qualquer outro site. Note que temos uma modelagem concisa que representa a rede de modo simples resta, então, definir uma métrica para representar a relevância de um site.

Pensando em Cadeias de Markov, e em sua evolução temporal, podemos pensar que uma boa métrica seja dada pela proporção do tempo que o internauta aleatório passa em cada site ao realizar um passeio aleatório pelos mesmos. Ou seja, vamos dizer que a relevância é maior para sites mais recorrentes e menor para os menos visitados. Dessa forma, podemos ver que sites que são muito linkados tendem a aparecer mais vezes no passeio. Consequentemente, os que são linkados pelos mais recorrentes também tendem a aparecer mais vezes. Já os sites pouco linkados e com links de sites menos recorrentes tendem a aparecer por menos tempo no passeio.

Note que essas propriedades são justamente as desejáveis para nosso ranking, ou seja, estamos com uma métrica em que os sites mais bem colocados tenham votos com peso maior que os piores colocados. Por fim, temos que a relevância pode ser calculada por meio do vetor estacionário da Cadeia de Markov definida pela rede.

Alguns problemas da modelagem

Vamos supor que uma sub-rede da nossa rede seja cíclica, como na Figura **??**. Se executarmos o passeio aleatório definido anteriormente temos que, se um internauta cair dentro desse ciclo, em A, por exemplo, ele ficará seguindo o caminho $A \to B \to C \to A$ infinitamente. Dessa forma, podemos ver que, ao entrar no ciclo, o tempo que o internauta permanece em cada site desse ciclo é $\frac{1}{3}$, o que nos daria igual relevância a cada um dos três sites do ciclo e relevância zero para os demais sites da rede, o que seria problemático.

Outro exemplo pode ser visto na Figura **??**, onde temos a aparição de um nó terminal. Note que, num passeio aleatório, se um internauta acessa o site G, ele permanece nele infinitamente, o seja, teremos relevância 1 para o site G e zero para os demais, o que também seria um problema.

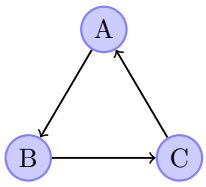

Figure 1: Exemplo de rede cíclica.

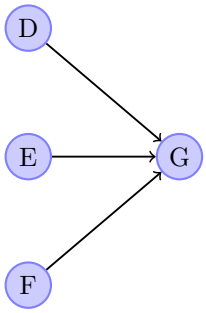

Figure 2: O nó G é um nó terminal.

Solucionando os problemas

Para driblar esses problemas podemos fazer algumas adaptações nas redes, alterando um pouco a estrutura e as probabilidades, mas visando uma maior coerência e correção desses problemas. A alteração estrutural se dará pela inserção de arestas de modo que o grafo fique completo, com todos os nós apontando para todos os outros nós. Quanto as probabilidades, fixaremos um valor $p \in (0, 1)$ e a matriz de transição dessa nova Cadeia será dada por

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

onde temos

- $\tilde{M}$: a matriz de transição da nova cadeia;
- M: a matriz de transição da cadeia original;
- I: a matriz de uns de tamanho $n \times n$ e;

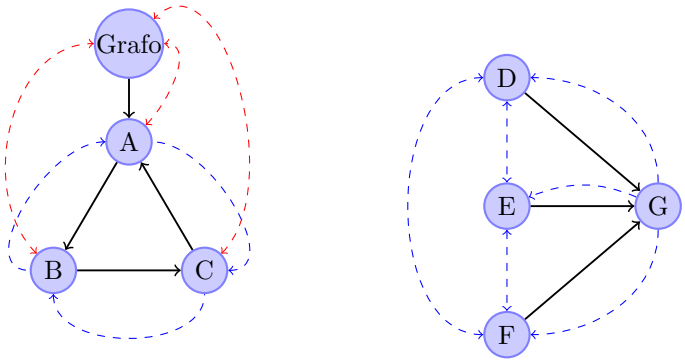

Figure 3: Grafos anteriores após alteração descrita.

- n: o número de nós da rede.

Visualmente, podemos ilustrar os grafos anteriores após as alterações como na Figura **??**. Na rede da esquerda, as arestas azuis representam as próprias arestas, enquanto as vermelhas representam grupos de arestas que se conectam a cada uma das arestas da outra parte do grafo, representada pelo no Grafo.

Uma interpretação dessa alteração é a de que o processo ocorre com o internauta, antes de sair do site, jogando uma moeda de probabilidade p para sair cara. Saindo cara, o internauta vai na barra de endereços e digita o endereço de qualquer site (podendo inclusive digitar o endereço do site atual) com mesma probabilidade, acessando o site que foi digitado na barra de endereços. Já se a moeda der coroa, então o internauta entra em um dos links da página, novamente com a mesma probabilidade entre os links existentes.

Podemos perceber que, com essas alterações, os problemas de ficarmos presos num ciclo ou num nó terminal acabam, pois temos uma probabilidade positiva de ir a outro nó da rede. Por fim, note também que essa alteração fez a matriz $\tilde{M}$ ser positiva o que, pelo Teorema de Perron-Frobenius, diz que essa matriz possui um estado estacionário π.

References

[1] Como nós começamos e onde estamos hoje. *Google*. https://about.google/our-story/.

[2] O algoritmo PageRank do Google. *Miguel Frasson - ICMC/USP*. https://edisciplinas.usp.br/pluginfile.php/5790758/mod_resource/content/1/pagerank-estat.pdf.

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