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

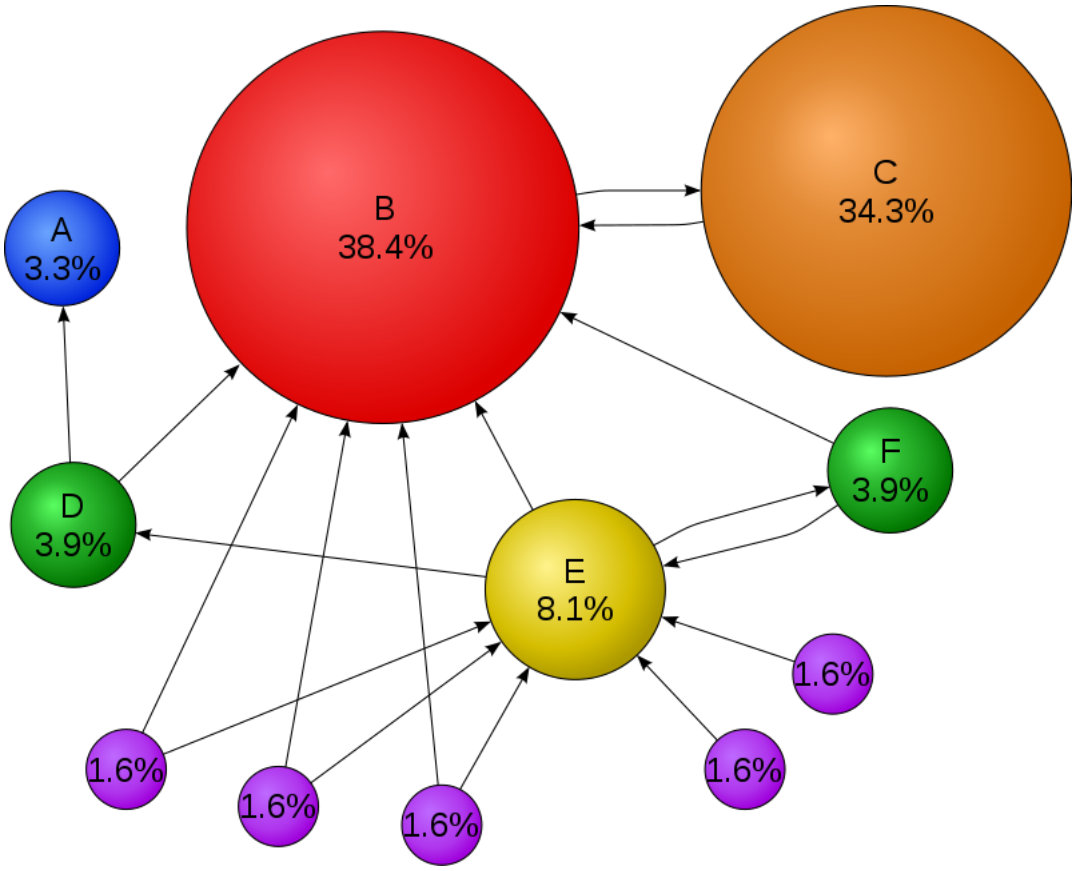
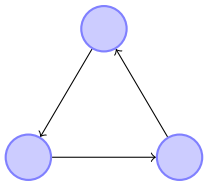


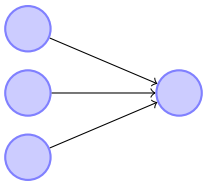
Figure 1: Páginas e suas relevâncias. Note que, mesmo com apenas um "voto", a página C tem uma alta relevância. Exemplo de [**?**].

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, com sua evolução temporal, podemos pensar que uma boa métrica seja dada pela proporção do tempo que o internauta aleatório passa em cada site, ou seja, é maior para sites mais recorrentes e menor para sites menos visitados. Note que essa definição é coerente com a ideia de que sites mais relevantes tenham votos mais significativos.

Tratando as entradas nulas e implementando





Como vimos anteriormente, podemos modelar as transições entre dois sites por meio de uma Cadeia de Markov. Porém, como sabemos, nem todos os sites recebem links de todos os outros sites (na verdade, a grande maioria nem se conecta), ou seja, nossa matriz de transição da cadeia é esparsa e pode não existir distribuição estacionária para essa cadeia. Uma forma de resolver esse problema é adicionar links fictícios entre todos os sites, mas de modo que esses links somem uma probabilidade baixa. Dessa forma, temos a seguinte matriz da cadeia

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

onde temos

- M: a matriz de transição da cadeia;
- p: a probabilidade de entrar em um site aleatório;
- A: a matriz de adjacências da rede;
- I: a matriz de uns de tamanho $n \times n$ e;
- n: o número de nós da rede.

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.