Page Rank e Cadeias de Markov

Cristhian Grundmann
Hanna Rodrigues Ferreira
Igor Cortes Junqueira
Igor Patrício Michels

December 1, 2021

Introdução

- Em 1995, Larry Page e Sergey Brin se conheceram na universidade de Stanford
- E objetivando criar um mecanismo de busca que pudesse

  "organizar as informações do mundo e torná-las universalmente acessíveis e úteis"

- Desenvolveram um algoritmo que usava os links para determinar a importância de cada página da internet [**?**].
- Ele ganhou força, virando o principal mecanismo de buscas
  - até então, as estratégias de calcular a relevância eram calculadas usando apenas os dados da própria página.

Cadeias de Markov

- Cadeias de Markov são um modelo estocástico que descreve uma sequência de eventos onde a probabilidade dos mesmos só depende do estado anterior.
- Processos de Markov são a base para simulações do tipo 'Monte Carlo Markov Chain'.
- podem ser representadas por uma 'matriz de transição', que descreve as probabilidades de transição de cada estado em particular para os demais possíveis estados.
- Também é possível obter a 'distribuição estacionária', que representa uma especie de 'equilíbrio' no processo a longo prazo.

Cadeias de Markov: O teorema de Perron-Frobenius

Seja A uma matriz $d \times d$ de entradas positivas, isto é $A_{ij} > 0 \ \forall i, j = 1, \ldots, d$. Então:

1. A possui um único autovetor x de norma 1, cujas componentes são, todas elas, positivas;
2. o autovalor λ_+ associado ao autovetor x é positivo e, para qualquer outro autovalor $\lambda \in \mathbb{C}$, temos que $|\lambda| < \lambda_+$;
3. o autovalor λ_+ é simples;

Page Rank

- O PageRank visa calcular a relevância de uma página por meio de fatores externos.
- podemos usar um grafo para ilustrar uma pequena rede, onde os nós são os sites e as arestas direcionadas representam 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?

Page Rank

- Em caso de empate, a ideia de Page e Brin foi ponderar os votos de acordo com a relevância
- 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 a Cadeia de Markov

- 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.
- Uma boa métrica para representar a relevância de um site seja dada pela proporção do tempo que o internauta passa em cada site ao realizar um passeio aleatório pelos mesmos.

Conexão com a Cadeia de Markov

- Ou seja, considerando que relevância é maior para sites mais recorrentes e menor para os menos visitados. conlcuímos que:
  1. sites que são muito linkados tendem a aparecer mais vezes no passeio.
  2. sites que são linkados pelos mais recorrentes também tendem a aparecer mais vezes.
  3. sites pouco linkados e com links de sites menos recorrentes tendem a aparecer por menos tempo no passeio.

- Logo, a relevância pode ser calculada por meio do vetor estacionário da Cadeia de Markov definida pela rede.

Alguns problemas na modelagem

- Vamos supor que uma sub-rede da nossa rede seja cíclica, ao executarmos o passeio aleatório e um internauta cair dentro dele, ele ficará no ciclo infinitamente.

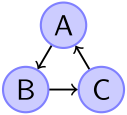

Figure: Exemplo de rede cíclica.

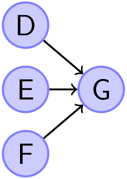

Figure: O nó G é um nó terminal.

- Quando temos a aparição de um nó terminal. No passeio aleatório, se o internauta acessar o site G, ele permanece nele infinitamente.

Solucionando os problemas

- podemos fazer algumas adaptações nas redes, alterando um pouco a estrutura e as probabilidades:
  - quanto a estrutura, faremos a inserção de arestas de modo que o grafo fique completo, com todos os nós apontando para todos os outros nós.
  - e quanto a 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;
    - n: o número de nós da rede.

Solucionando os problemas

- 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 no Grafo.

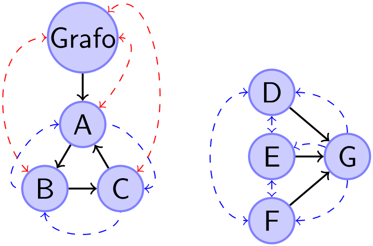

Figure: Grafos anteriores após alteração descrita.

Exemplo

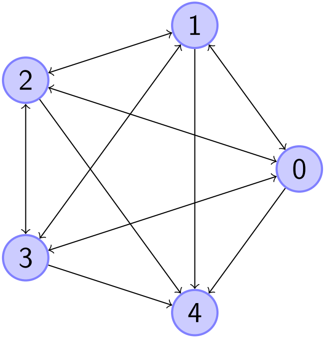

Figure: Exemplo de rede.

Referências I

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

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

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

Referências II

O Teorema de Perron-Frobenius e a Ausência de Transição de Fase em Modelos Unidimensionais da Mecânica Estatíıstica. *Marcelo Richard Hilário - UFMG. Gastão Braga - UFMG*. `https://www.ime.usp.br/~map2121/2014/map2121/programas/perron-frobenius.pdf`.