PageRank do Google

Álgebra Linear

Alunos: Ricael Daniel Oliveira da Silva e
Thiago Franke Melchiors
Professor: Yuri Fahham Saporito

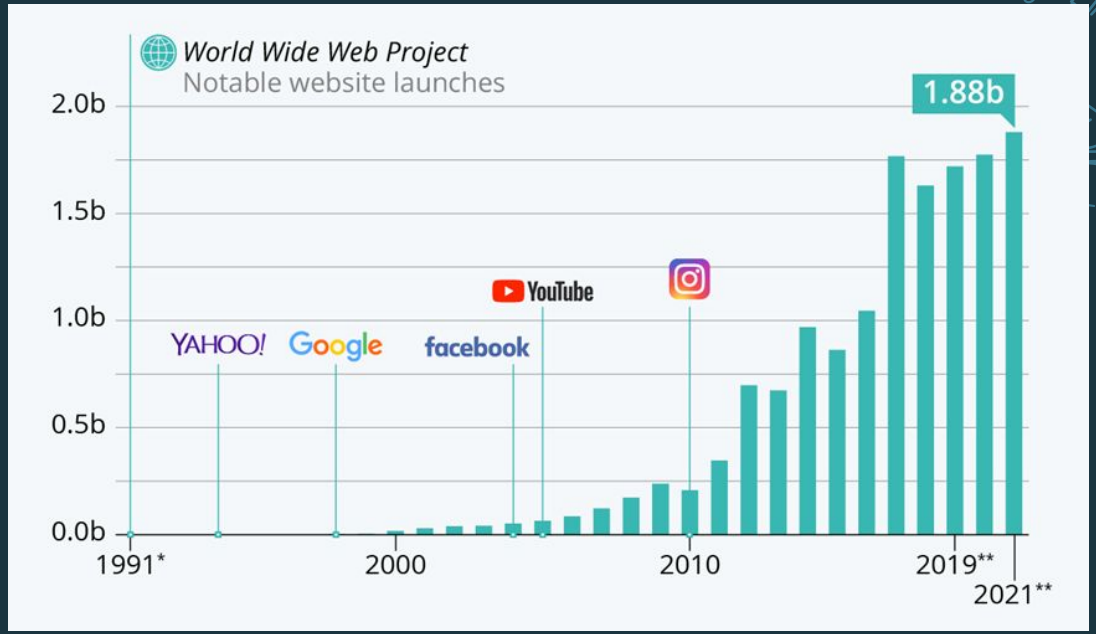
Sumário

1. Contexto Histórico

1.1. A dificuldade no gerenciamento de páginas na internet

A organização, classificação e gerenciamento dos sites tornou-se mais difícil na mesma medida que as páginas de web se multiplicavam.

Quantos Websites existem?



1. Contexto Histórico

1.1. A dificuldade no gerenciamento de páginas na internet

Os primeiros motores de busca da internet possuíam limitações:

- Combinações das palavras-chave nas pesquisas não eram bem interpretadas;
- Os resultados das consultas eram relacionados a publicidade de produtos.

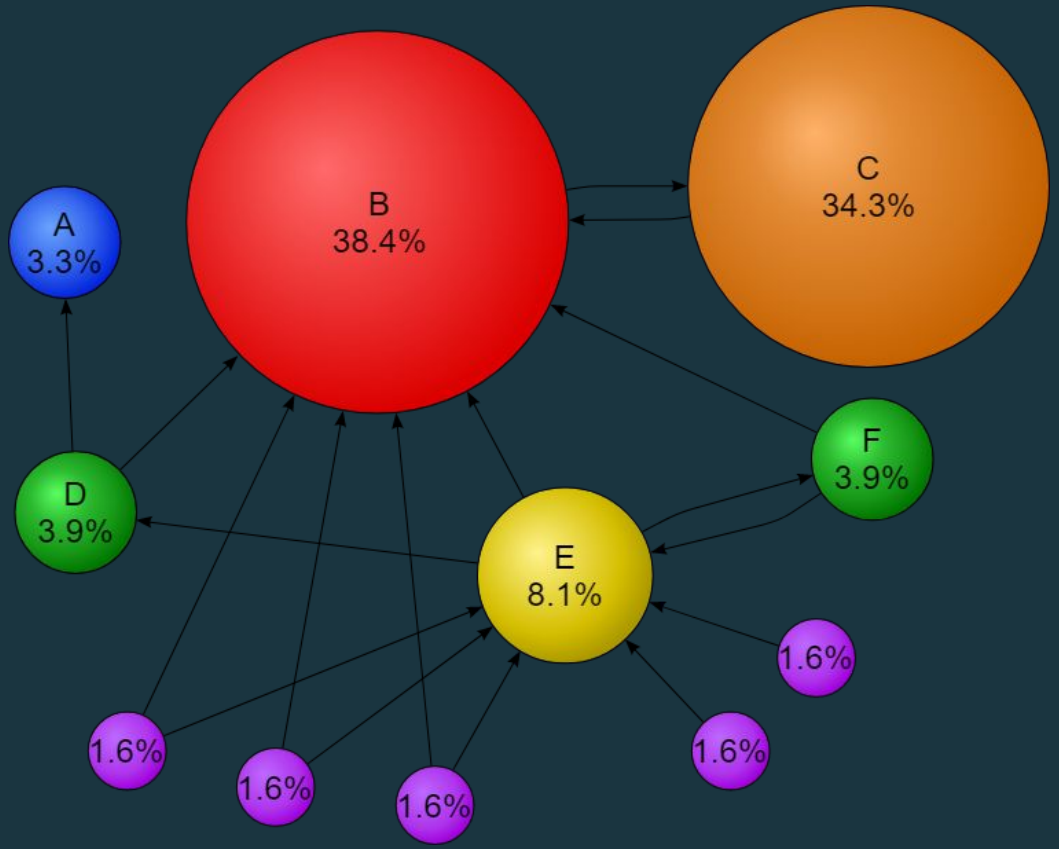
Em 1998, Larry Page e Sergey Brin desenvolveram o **PageRank**:

- Algoritmo de classificação da importância dos websites com base no número de referências entre as páginas.

Larry e Sergey aplicaram o algoritmo na empresa que criaram, a Google Inc.

1. Contexto Histórico

1.2. Uma Breve apresentação do mecanismo



2. Funcionamento do algoritmo

2.1. Esclarecimentos

01

O PR de cada página representa a probabilidade de um usuário, navegando aleatoriamente, chegar até ela. A soma de todos os ranqueamentos é igual a 1.

02

O algoritmo é iterativo.

03

Após a primeira iteração, todas as páginas terão a classificação 1/N (N é o número total de páginas).

2. Funcionamento do algoritmo

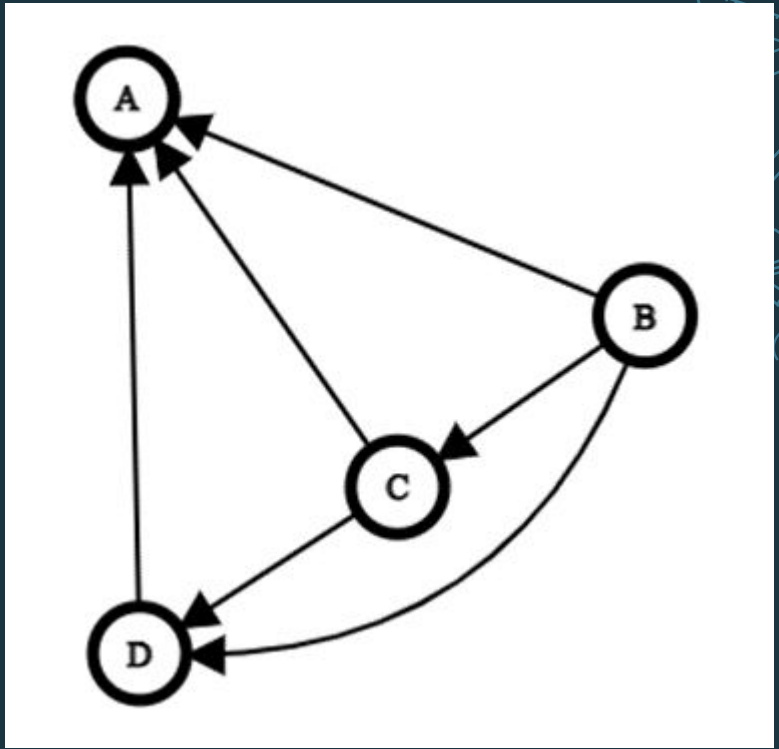
2.2. O algoritmo simplificado

Para encontrar o PageRank do nó A:

1º Após a primeira iteração, as páginas têm o mesmo valor 1/N de PageRank, ou seja, 0,25.

2º Na segunda iteração, cada página transfere o seu PageRank em porções iguais para as páginas que aponta.

$$PR(A) = \frac{PR(B)}{3} + \frac{PR(C)}{2} + \frac{PR(D)}{1}$$



2. Funcionamento do algoritmo

2.2. O algoritmo simplificado

Equação geral:

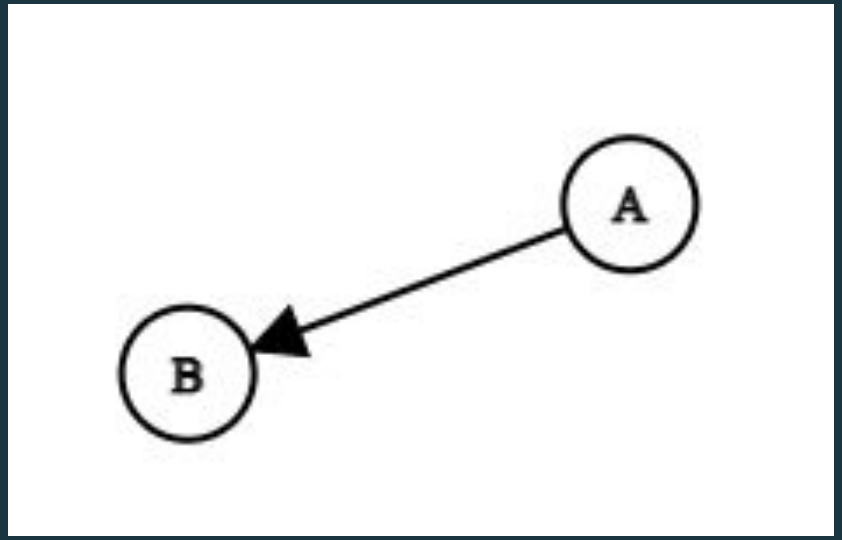
$$PR(p_i) = \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

M(pi) conjunto de todas as páginas pj que referenciam a página pi;
L(pj) número de referências em pj.

3. Falhas do PageRank

3.1. Páginas sem ligações

B absorve o ranqueamento da rede, mas ao calcular o PageRank de cada página, ambos os nós recebem zero.



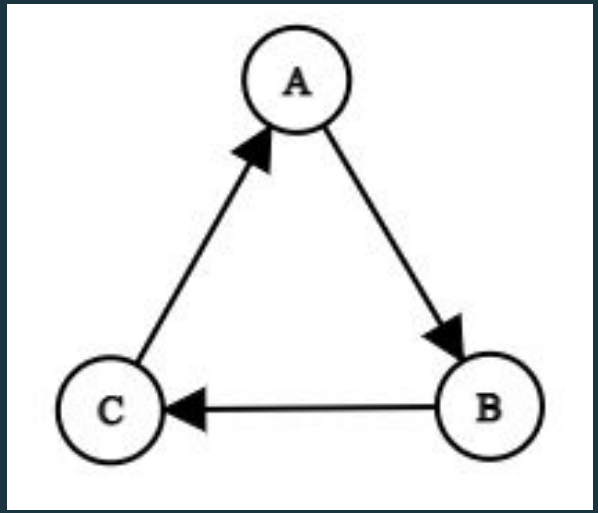
3. Falhas do PageRank

3.2. Ciclos (Rank Sink)

Após cada iteração, o valor de ranqueamento é transferido para o nó seguinte em sua totalidade.

Não há um momento em que o sistema entra em equilíbrio.

O cálculo resulta em 1 de PageRank para todas as redes: ABSURDO!



3. Falhas do PageRank

3.3. Solução: Fator de Amortecimento

- Representado por d;
- Varia de 0 a 1;
- Probabilidade de um usuário continuar seguindo as ligações entre as páginas;
- O valor padrão para d é 0, 85;

- Se d = 0, todas as páginas ficam com o PageRank de 1/N;
- Quanto mais próximo d estiver de 1, maior é a influência da estrutura da rede;

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}\right)$$

4. Representação Matricial

$$PR(p_i) = \frac{1-d}{N} + d \left(\sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \right)$$

$$R = \begin{bmatrix} \frac{1-d}{N} \\ \frac{1-d}{N} \\ \vdots \\ \frac{1-d}{N} \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & \cdots & l(p_1,p_n) \\ l(p_2,p_1) & \ddots & & \vdots \\ \vdots & & l(p_i,p_j) & \\ l(p_n,p_1) & \cdots & & l(p_n,p_n) \end{bmatrix} R$$

Exemplo computacional

Como seria a classificação da premier league temporada 20/21 se a classificação fosse por PageRank em vez de pontos corridos?

Base de dados original:

```
In [4]: data.head(11)
```

Out[4]:

	Match Number	Round Number	Date	Location	Home Team	Away Team	Result
0	4	1	12/09/2020 12:30	Craven Cottage	Fulham	Arsenal	0 - 3
1	3	1	12/09/2020 15:00	Selhurst Park	Crystal Palace	Southampton	1 - 0
2	5	1	12/09/2020 17:30	Anfield	Liverpool	Leeds	4 - 3
3	8	1	12/09/2020 20:00	London Stadium	West Ham	Newcastle	0 - 2
4	7	1	13/09/2020 14:00	The Hawthorns	West Brom	Leicester	0 - 3
5	6	1	13/09/2020 16:30	Tottenham Hotspur Stadium	Spurs	Everton	0 - 1
6	10	1	14/09/2020 18:00	Bramall Lane	Sheffield Utd	Wolves	0 - 2
7	9	1	14/09/2020 20:15	Amex Stadium	Brighton	Chelsea	1 - 3
8	14	2	19/09/2020 12:30	Goodison Park	Everton	West Brom	5 - 2
9	15	2	19/09/2020 15:00	Elland Road	Leeds	Fulham	4 - 3
10	17	2	19/09/2020 17:30	Old Trafford	Man Utd	Crystal Palace	1 - 3

Alterando a base de dados original:

```
In [45]: df.head(11)
```

Out[45]:

	home_team	away_team	home_gols	away_gols
0	Fulham	Arsenal	0	3
1	Crystal Palace	Southampton	1	0
2	Liverpool	Leeds	4	3
3	West Ham	Newcastle	0	2
4	West Brom	Leicester	0	3
5	Spurs	Everton	0	1
6	Sheffield Utd	Wolves	0	2
7	Brighton	Chelsea	1	3
8	Everton	West Brom	5	2
9	Leeds	Fulham	4	3
10	Man Utd	Crystal Palace	1	3

Criando a matriz de transição:

```
In [25]: df_matches.head()
```

Out[25]:

	winner	loser
0	Arsenal	Fulham
1	Crystal Palace	Southampton
2	Liverpool	Leeds
3	Newcastle	West Ham
4	Leicester	West Brom

```
In [15]: matrix.head()[matrix.columns[:5]]
```

Out[15]:

	Arsenal	Aston Villa	Brighton	Burnley	Chelsea
Arsenal	NaN	NaN	NaN	NaN	NaN
Aston Villa	NaN	NaN	NaN	NaN	NaN
Brighton	NaN	NaN	NaN	NaN	NaN
Burnley	NaN	NaN	NaN	NaN	NaN
Chelsea	NaN	NaN	NaN	NaN	NaN

Criando a matriz de transição:

In [41]:
```
# Matriz de transição M.
M.head()[M.columns[:5]]
```

Out[41]:

	Arsenal	Aston Villa	Brighton	Burnley	Chelsea
Arsenal	0.0	0.0	1.0	1.0	1.0
Aston Villa	1.0	0.0	1.0	1.0	1.0
Brighton	0.0	1.0	0.0	1.0	1.0
Burnley	1.0	1.0	1.0	0.0	0.0
Chelsea	0.0	1.0	1.0	1.0	0.0

Finalizando a matriz de transição:

```
In [46]: dfpr.head()[dfpr.columns[:5]]
```

Out[46]:

	Arsenal	Aston Villa	Brighton	Burnley	Chelsea
Arsenal	0.000000	0.000000	0.058824	0.058824	0.076923
Aston Villa	0.071429	0.000000	0.058824	0.058824	0.076923
Brighton	0.000000	0.058824	0.000000	0.058824	0.076923
Burnley	0.071429	0.058824	0.058824	0.000000	0.000000
Chelsea	0.000000	0.058824	0.058824	0.058824	0.000000

Após finalizarmos a matriz de transição, calculamos o pagerank utilizando a fórmula abaixo.

$$Pr = \frac{1-d}{n} + d \cdot L$$

Após isso calculamos os autovalores e autovetores da matriz Pr, e então pegamos o autovetor correspondente ao autovalor 1 para ser o pagerank dos times.

In [40]: classificacao

Out[40]:

	Time	Pagerank
1	Liverpool	0.273477
2	Man Utd	0.272085
3	Man City	0.266215
4	Leicester	0.262372
5	Chelsea	0.261980
6	Spurs	0.258811
7	Everton	0.243673
8	Leeds	0.238636
9	Brighton	0.234189
10	Aston Villa	0.220853
11	Crystal Palace	0.212731
12	West Ham	0.212430
13	Southampton	0.206356
14	Fulham	0.202996
15	Arsenal	0.201735
16	West Brom	0.187869
17	Wolves	0.182869
18	Newcastle	0.180655
19	Burnley	0.156745
20	Sheffield Utd	0.122887

POS	CLUB	MP	GD	PTS
1	Manchester City	38	51	86
2	Manchester United	38	29	74
3	Liverpool	38	26	69
4	Chelsea	38	22	67
5	Leicester City	38	18	66
6	West Ham	38	15	65
7	Tottenham Hotspur	38	23	62
8	Arsenal	38	16	61
9	Leeds United	38	8	59
10	Everton	38	-1	59
11	Aston Villa	38	9	55
12	Newcastle United	38	-16	45
13	Wolverhampton Wanderers	38	-16	45
14	Crystal Palace	38	-25	44
15	Southampton	38	-21	43
16	Brighton Hove & Albion	38	-6	41
17	Burnley	38	-22	39
18	Fulham	38	-26	28
19	West Bromwich Albion	38	-41	26
20	Sheffield United	38	-43	23

CHAMPIONS LEAGUE GROUP STAGE EUROPA LEAGUE GROUP STAGE RELEGATION

Resultados interessantes:

- O Liverpool teria sido o campeão.
- O Fulham sairia da 18° para a 14° colocação.
- O West Brom sairia da 19° para a 16° colocação.
- O Burnley sairia de 17° para 19° e seria rebaixado.
- O Newcastle sairia de 12° para 18° e seria rebaixado .