KGŠM

Ako Google určuje dôležitosť webstránok?

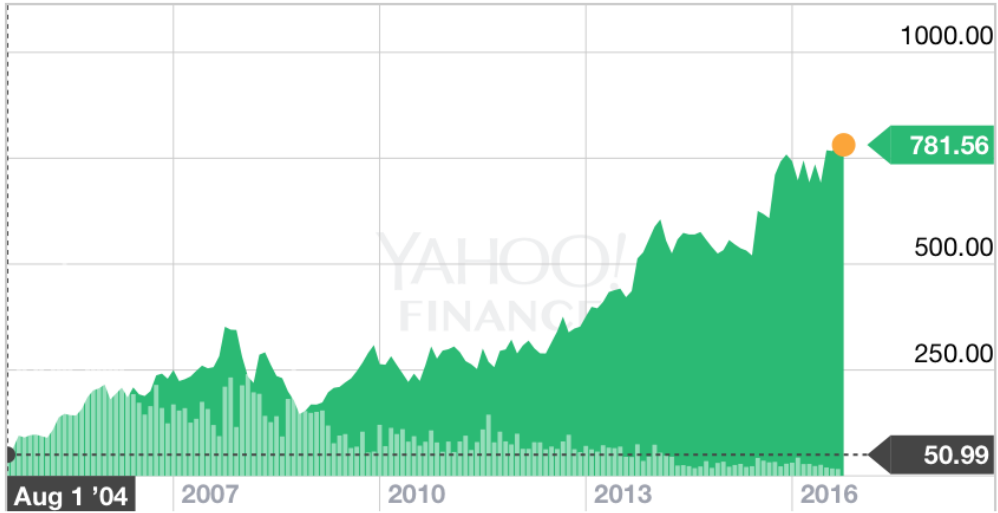
Lukáš Lafférs

KM FPV UMB
www.lukaslaffers.com

14. November, 2022

www.google.com

1D 5D 1M 6M YTD 1Y 2Y 5Y 10Y **MAX** Interactive chart

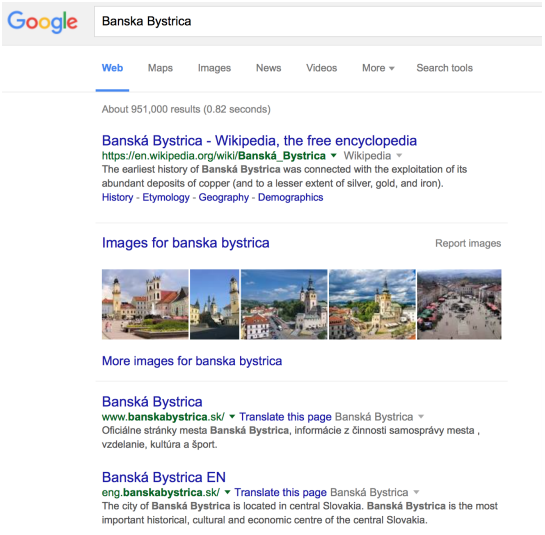


Larry Page a Sergey Brin

Larry Page a Sergey Brin

- PhD študenti na Stanfordskej Univerzite
- 1999 - algoritmus PageRank
- patent $\rightarrow$ \$1,8 mil. $\rightarrow$ \$336 mil.

Príklad


Google
Banska Bystrica
Web
Maps
Images
News
Videos
More
Search tools
About 951,000 results (0.82 seconds)
Banská Bystrica - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Banská_Bystrica Wikipedia
The earliest history of Banská Bystrica was connected with the exploitation of its abundant deposits of copper (and to a lesser extent of silver, gold, and iron).
History - Etymology - Geography - Demographics
Images for banska bystrica
Report images
More images for banska bystrica
Banská Bystrica
www.banskabystrica.sk/ Translate this page Banská Bystrica
Oficiálne stránky mesta Banská Bystrica, informácie z činnosti samosprávy mesta , vzdelanie, kultúra a šport.
Banská Bystrica EN
eng.banskabystrica.sk/ Translate this page Banská Bystrica
The city of Banská Bystrica is located in central Slovakia. Banská Bystrica is the most important historical, cultural and economic centre of the central Slovakia.

Problém

Chceme usporiadať stránky podľa dôležitosti.

Ako ?

Čo majú dôležité stránky spoločné?

- Existujú stránky, ktoré na ne odkazujú.
- Existuje **veľa** stránok, ktoré na ne odkazujú.
- Existuje veľa **dôležitých** stránok, ktoré na ne odkazujú.

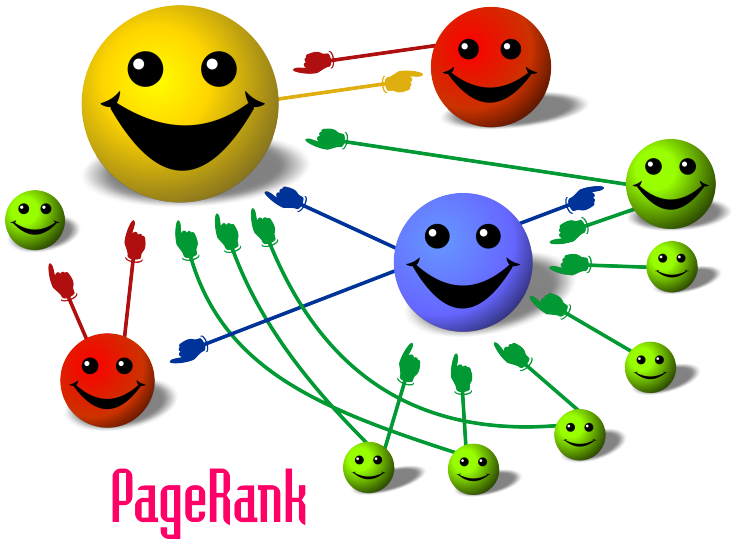
Príklady dôležitých stránok:

- `www.wikipedia.org`
- `www.google.com`
- `www.twitter.com`
- `www.facebook.com`
- `www.nasa.gov`
- `www.microsoft.com`

Príklady menej dôležitých stránok:

- `www.akozasaditcibulu.sk`
- `www.akozasaditcesnak.sk`
- `www.varimecibulu.sk`
- `www.varimecviklu.sk`
- `www.akoneprevaritcviklu.sk`
- `www.lukaslaffers.com`

Web ako sieť stránok



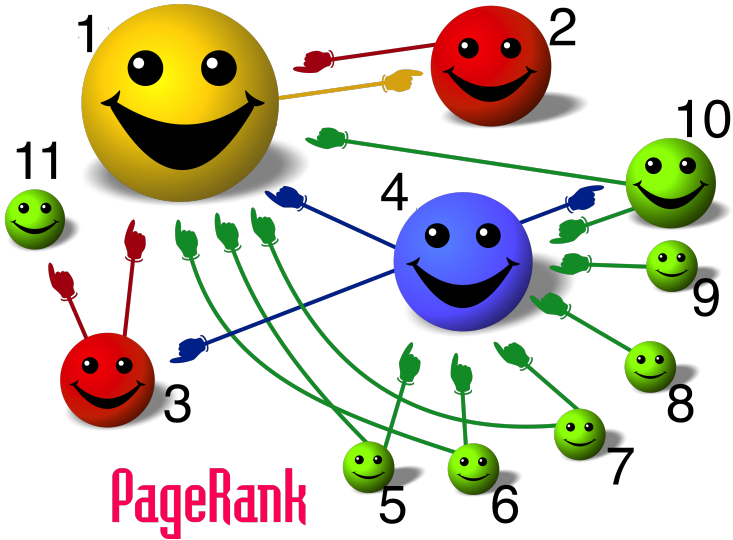
wikimedia commons

Problém náhodného surfujúceho

Náhodný surfujúci prišiel na nejakú webstránku, čo urobí?

- S pravdepodobnosťou 85% šťukne na náhodný odkaz
- S pravdepodobnosťou 15% zatvorí browser a ide na úplne inú, náhodnú stránku

Náhodný surfujúci si nič nepamätá, len šťuká a šťuká...


1
2
10
11
4
9
3
8
7
5
6
PageRank

Pravdepodobnosti prechodu medzi stránkami

$$M = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Zavriem browser a šťuknem na náhodnú stránku

$$N = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

Náhodne šťukajúci surfer

$$0.85M + 0.15N$$

Náhodne šťukajúci surfer

$$P = \begin{pmatrix}
0.015 & 0.440 & 0.015 & 0.440 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.865 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.865 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.298 & 0.015 & 0.298 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.298 \\
0.440 & 0.015 & 0.015 & 0.440 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.440 & 0.015 & 0.015 & 0.440 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.440 & 0.015 & 0.015 & 0.440 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.015 & 0.015 & 0.015 & 0.865 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.015 & 0.015 & 0.015 & 0.865 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\
0.440 & 0.015 & 0.015 & 0.440 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015
\end{pmatrix}$$

0.85 * 

0.15 * 

Koľko percent času strávi náhodne šťukajúci surfer na nejakej webstránke?

Čím viacej, tým je stránka dôležitejšia!

Náhodne šťuká, na ktorej stránke bol najviackrát?

alebo

Veľa surfujúcich. Na ktorej stránke ich je najviac?

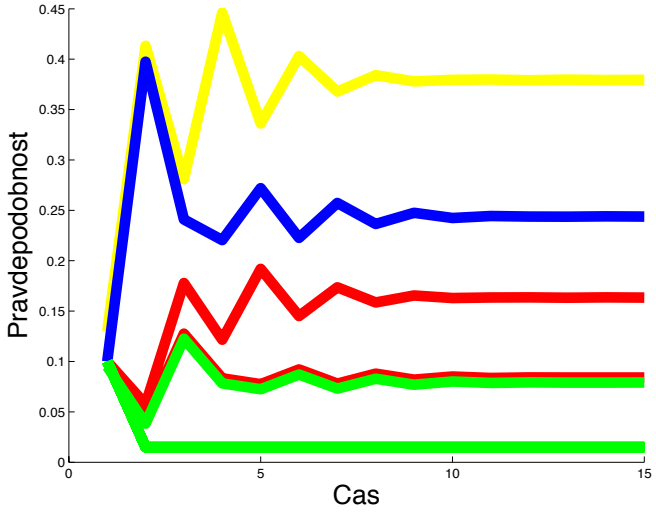
Veľa surfujúcich

Na ktorej stránke začnú?

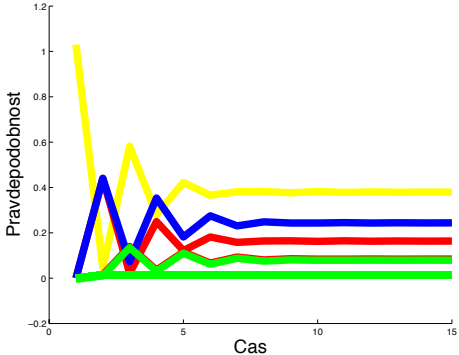
Napr.:

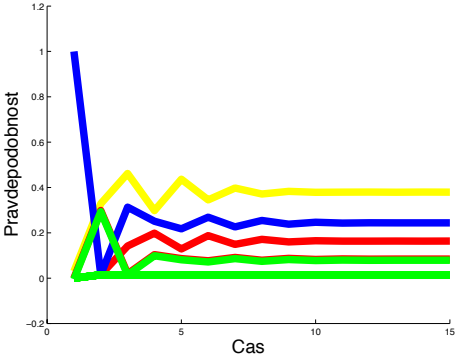
- Na každej stránke rovnako veľa.
- Všetci začnú na tej istej stránke.

Rozdelím ich spravodlivo

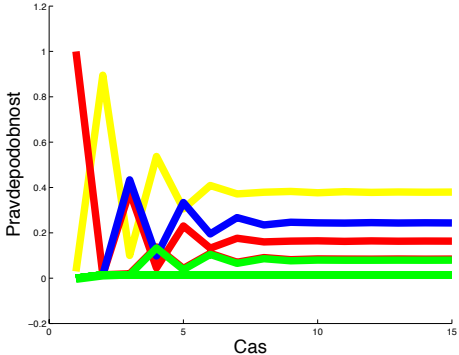


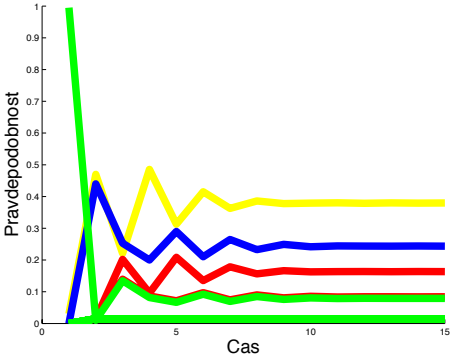
Rozdelím ich nespravodlivo (1)

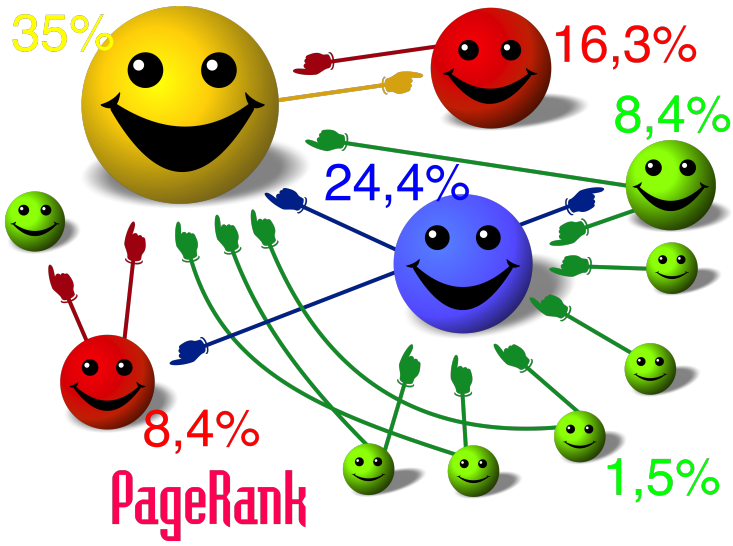




Rozdelím ich nespravodlivo (2)






35%
16,3%
8,4%
24,4%
8,4%
1,5%
PageRank

Nech ich rozdelíme akokoľvek, za nejaký čas to bude vyzerať vždy rovnako.

Toto rozdelenie sa nazýva **stacionárne**.

pagerank.m

```
1   % PageRank example
2
3   % zostrojime maticu susednosti
4   clear all
5
6   Susednosti = zeros(10,10);
7   Susednosti(1,2) = 1;
8   Susednosti(1,4) = 1;
9   Susednosti(2,1) = 1;
10  Susednosti(3,1) = 1;
11  Susednosti(4,[1 3 10]) = [1 1 1];
12  Susednosti(5,[1 4]) = [1 1];
13  Susednosti(6,[1 4]) = [1 1];
14  Susednosti(7,[1 4]) = [1 1];
15  Susednosti(8,4) = 1;
16  Susednosti(9,4) = 1;
17  Susednosti(10,[1 4]) = [1 1];
18
19  n = 10;
20
21  MaticaPrechodu =...
22      Susednosti.*...
23      (1./sum(Susednosti'))'*...
24      ones(1,n));
25
26  MaticaZmeny = (1/n)*ones(n,1)*ones(1,n);
27
28  pravdZmeny = 0.15;
29
30  M = MaticaZmeny*pravdZmeny + MaticaPrechodu*(1-pravdZmeny);
31
32  MM = M^100;
33
34  stacionarnaDistribucia = MM(1,:);
35
```

Nie je to až tak jednoduché

Naš príklad: 10 webstránok

Google: 130 000 000 000 webstránok (Nov 2016).

- To ako Google usporiadava stránky je v skutočnosti extrémne tajná a cenná informácia
- Ich algoritmus používa niekoľko stoviek typov informácií na to, aby odporučil tú najlepšiu stránku
- Každý rok implementujú okolo 500 zmien a urobia viac ako 20000 experimentov
- TrustRank - dôveryhodnejšie stránky sú dôležitejšie
- Google je dobrý v odhaľovaní toho, kto sa snaží okabátiť PageRank
- Je oveľa ťažsie zvýšiť $PR7 \rightarrow PR8$ ako $PR1 \rightarrow PR2$

Stacionárne rozdelenie je v prípade Google riešenie veľa rovníc o veľa neznámych.

Znalosťami z algebry však vieme, že hľadáme tzv. **vlastný vektor** a na to poznáme sofistikované spôsoby ako to vyrátať aj pre obrovské množstvo webstránok.

$$\pi = \pi P$$

$$P = \begin{bmatrix} \text{  } & \text{  } \\ \text{  } & \text{  } \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

Dnes je 

Aké počasie bude o polroka?

$$\pi = \pi \begin{bmatrix} \text{  } & \text{  } \\ \text{  } & \text{  } \end{bmatrix}$$

$$\pi = [0.714, 0.285]$$

$$Pr(\text{  }) = 71.4\%$$

$$Pr(\text{  }) = 28.5\%$$

$$P = \begin{bmatrix} \text{  } & \text{  } \\ \text{  } & \text{  } \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}$$

Dnes je 

Aké počasie bude o polroka?

$$Pr\left(\text{  }\right) = 99.999999....\%$$

$$Pr\left(\text{  }\right) = 0.0000....1\%$$

Náhodný proces surfujúceho sa volá **Markovov reťazec** s prechodovými pravdepodobnosťami danými tabuľkou P.

- Modelovanie rizikovosti finančných produktov
- Správanie sa dynamiky plynov
- Modelovanie aktivity enzýmov
- Rozoznávanie reči
- Optimalizácia telekomunikačných sietí
- Makroekonomické modelovanie
- Genetika
- Počasie
- Predikovanie víťazstva v bejzbale

Zhrnutie

- Larry Page a Sergey Brin sa v roku 1996 zaoberali otázkou ako zoradiť webstránky podľa dôležitosti
- Nápad náhodného surfujúceho im umožnil jednoduchú matematickú formuláciu problému
- Na riešenie tohoto problému potom vedeli použiť efektívne algoritmy
- ich myšlienka zmenila spôsob akým dnes vyhľadávame informácie

Ďakujem za pozornosť!