Motores de pesquisa na Internet

- Pessoas introduzem palavras numa textbox.

- Motor de pesquisa devolve páginas contendo essas palavras.

Porque é que é díficil?

- 250 milhões de pesquisas por dia
  (6000 por segundo em horas de ponta!)

- Cada pessoa quer obter respostas em menos de 0.2 sec.

Panorâmica geral

1. Obter uma cópia da Web.

2. Construir um índice da Web por palavra.

3. Decidir quais as páginas que aparecem nos primeiros lugares.

Obter uma cópia da Web (crawling)

- Um webcrawler começa por ir buscar uma página.
- Segue links que saem dessa página.
- ... continua até uma determinada condição de paragem.
- condição de paragem no Google = 4.3 biliões de páginas.
- Guarda o texto da Web em disco.

Web crawling: parece fácil mas não é

- Muitos sites são infinitos.
- Muitos sites podem estar temporáriamente em baixo.
- Tem de se ter cuidado em não "crashar" sites.

Web crawling

- Utilizar muitos crawlers.
- Google: centenas de computadores ligados a uma linha T3 (45 Mbits/sec.)
- Por onde é que os crawlers começam?

Construir um índice da Web por palavra

- Um índice associa palavras a páginas.
- Guardar informação auxiliar para cada ocorrência de cada palavra.
- Muita informação. Actualmente, cerca de 10 terabytes para um índice de 4.3 biliões de páginas.

Depois de ter o índice...

- Pessoas introduzem palavras numa textbox.

- Motor de pesquisa devolve páginas contendo essas palavras.

Google

- Fora com operadores booleanos — 99.9% das pessoas querem fazer AND.

- Quais as páginas que devem aparecer nos primeiros lugares?

- Solução do Google: PageRank.

PageRank

- PageRank é um algoritmo que ordena as páginas da Web por importância.

- Inspirado no *Science Citation Index*.

- Um link para uma página dá importância a essa página.

- ...mas só isso não chega. Porquê?

PageRank (cont.)

- A minha página é importante se páginas importantes apontarem para mim.

- A importância de uma página é a probabilidade que essa página tem em ser visitada por um surfista aleatório.

PageRank (cont.)

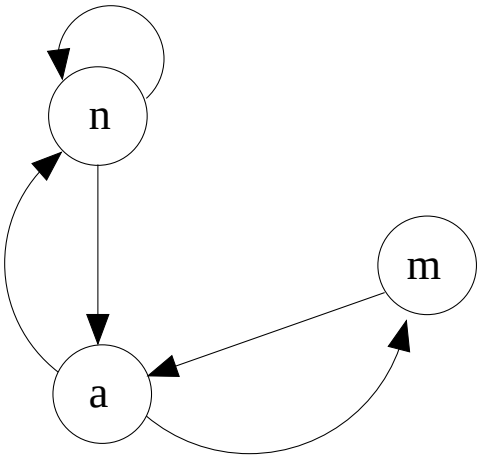
- Criar um grafo da Web (páginas são nós, links são arcos).

- Inicialmente cada página tem 1 unidade de importância.

- Em cada iteração, cada página dá importância aos seus sucessores e recebe importância dos seus antecessores.

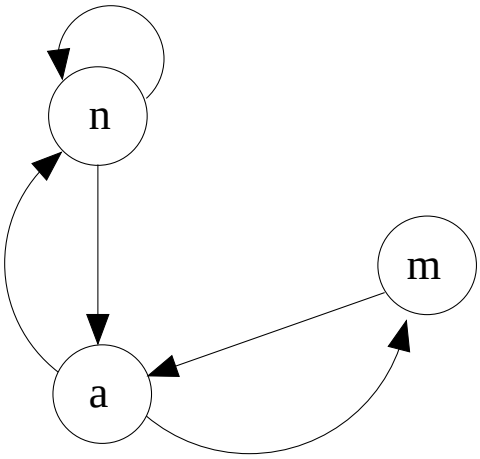
- Calcular a distribuição estacionária de um surfista aleatório.

A Web com 3 sites: Netscape, Amazon, Microsoft



$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

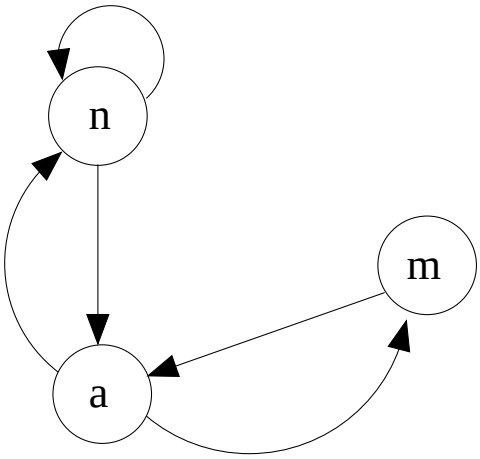
A Web com 3 sites: Netscape, Amazon, Microsoft



$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- Inicialmente: $n = m = a = 1$

A Web com 3 sites: Netscape, Amazon, Microsoft

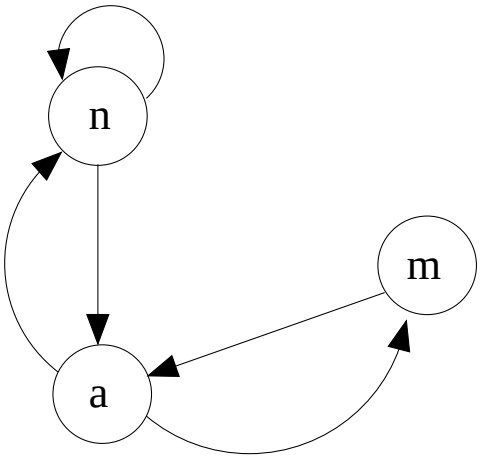


$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- Inicialmente: $n = m = a = 1$

- Na 2ª iteração: $n = 1$, $m = 1/2$, $a = 3/2$

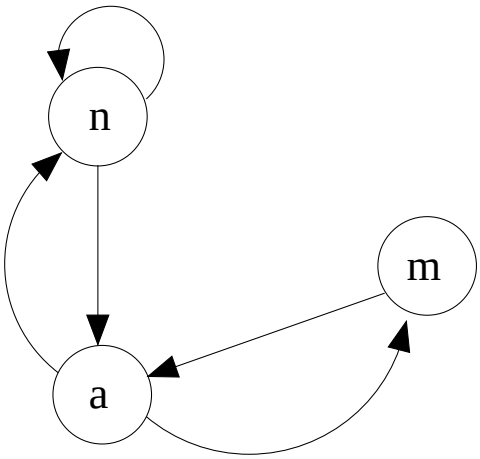
A Web com 3 sites: Netscape, Amazon, Microsoft



$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- Inicialmente: $n = m = a = 1$
- Na 2ª iteração: $n = 1$, $m = 1/2$, $a = 3/2$
- Na 3ª iteração: $n = 5/4$, $m = 3/4$, $a = 1$

A Web com 3 sites: Netscape, Amazon, Microsoft



$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- Inicialmente: $n = m = a = 1$
- Na 2ª iteração: $n = 1$, $m = 1/2$, $a = 3/2$
- Na 3ª iteração: $n = 5/4$, $m = 3/4$, $a = 1$
- ∞ iterações: $n = 6/5$, $m = 3/5$, $a = 6/5$

PageRank

- PageRank foi inventado por Larry Page, um dos fundadores da empresa.

- É difícil de enganar o PageRank.

- Porque é que um esquema baseado unicamente no conteúdo de uma página falha?

Robustez e eficiência

- Indice e Web divididos em bocados.
- Cada bocado é replicado 40-60 vezes.
- Um front-end de servidores Web (cerca de 500) ligados a cópias redundantes da Web.
- Load balancing no front-end.
- Hardware simples e barato, a correr Linux.
- 10000 computadores.

Linux porquê?

1. $$

2. Open source $\rightarrow$ 64-bit filesytem

3. ...

Linguagens de programação

1. Crawler $\rightarrow$ python

2. O resto $\rightarrow$ C e C++

Simplicidade

- O cliente tem sempre razão.
- Fora com banners e mariquices a saltitar!
- User-interface é muito simples —
  "It's hard to mess up a simple page".

Mais coisas sobre o Google

- Actualmente tem cerca de 2000 empregados.

- Os donos da empresa, Larry Page e Sergey Brin, tinham 25 e 26 anos quando fundaram o google.

- Eram alunos na Universidade de Stanford.