A matemática por trás do Google

Seminário de Coisas Legais - 16/08/2012

Tiago J. Fonseca

Background histórico

- Em 1990 a internet é criada dentro do CERN, e rapidamente começa a se tornar popular no meio acadêmico.

Background histórico

- Em 1990 a internet é criada dentro do CERN, e rapidamente começa a se tornar popular no meio acadêmico.

Background histórico

- Em 1990 a internet é criada dentro do CERN, e rapidamente começa a se tornar popular no meio acadêmico.

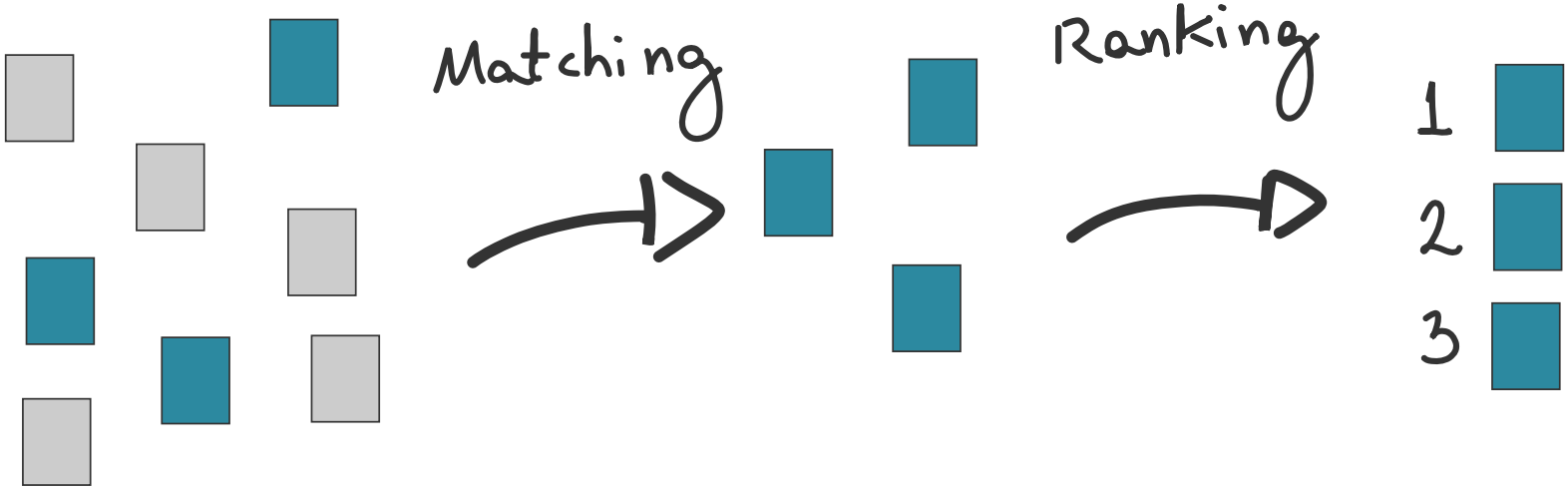
- Em 1993 surgiram os primeiros “buscadores”.

O que é um buscador?

- Sabemos o que um buscador faz, mas como ele funciona?

O que é um buscador?

- Sabemos o que um buscador faz, mas como ele funciona?
- Essencialmente, todo buscador desempenha 2 passos:
  1. **Matching**: O algoritmo busca, dentre todas as páginas da Web, aquelas que contêm as palavras (ou frases) digitadas.
  2. **Ranking**: O algoritmo seleciona quais, dentre as páginas encontradas no passo 1, são as mais "relevantes"e ordena o resultado.
- O passo mais decisivo é o *ranking*!



Continuando...

- A internet continuou se popularizando e diversos sites de busca foram criados.

Continuando...

- A internet continuou se popularizando e diversos sites de busca foram criados.
- Cada um, tentava trazer alguma inovação em seu algoritmo de *ranking*.

Continuando...

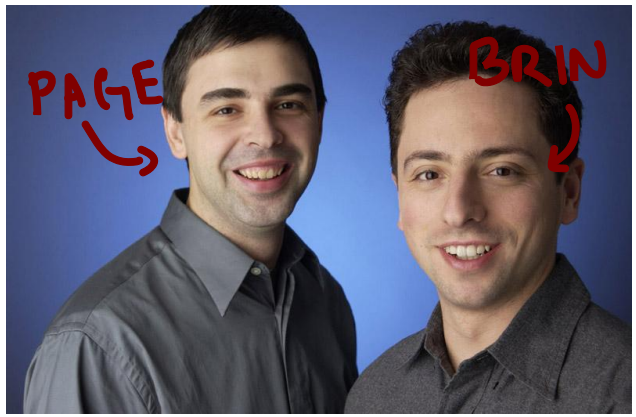
- A internet continuou se popularizando e diversos sites de busca foram criados.
- Cada um, tentava trazer alguma inovação em seu algoritmo de *ranking*.
- Entre 1994 e 1998, os buscadores mais famosos eram o Lycos e o AltaVista (que já não existe mais).

Google!

- Em 1997 e 1998 surge um novo buscador, o Google.

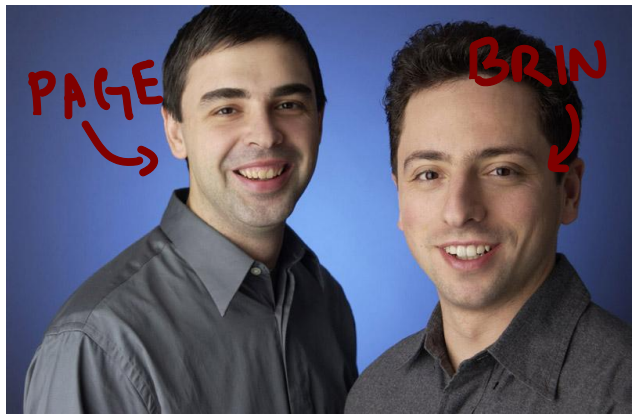
Google!

- Em 1997 e 1998 surge um novo buscador, o Google.
- Criado pelos doutorandos em Ciência da Computação, em Stanford, Larry Page and Sergey Brin.



Google!

- Em 1997 e 1998 surge um novo buscador, o Google.
- Criado pelos doutorandos em Ciência da Computação, em Stanford, Larry Page and Sergey Brin.



- O Google tinha algo de diferente, pouco depois do seu lançamento, a revista *PC Magazine* o elegeu como um dos 100 melhores sites da época e comentou que o Google possuia "*an uncanny knack for returning extremely relevant results*".

Google!

- Em 1997 e 1998 surge um novo buscador, o Google.
- Criado pelos doutorandos em Ciência da Computação, em Stanford, Larry Page and Sergey Brin.

- O Google tinha algo de diferente, pouco depois do seu lançamento, a revista *PC Magazine* o elegeu como um dos 100 melhores sites da época e comentou que o Google possuia "*an uncanny knack for returning extremely relevant results*".
- Boa parte desse sucesso foi devido às inovações no seu algoritmo de *ranking*, o **PageRank**.

Primeira tela do Google

Primeira tela do Google

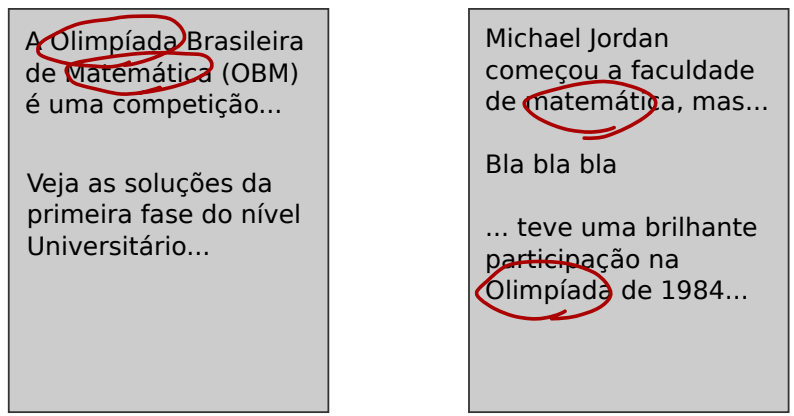
Um pouco sobre Ranking

Um pouco sobre Ranking

- Em geral, as técnicas de ranking usadas pelos algoritmos de *ranking* anteriores eram focadas no **conteúdo das páginas**.

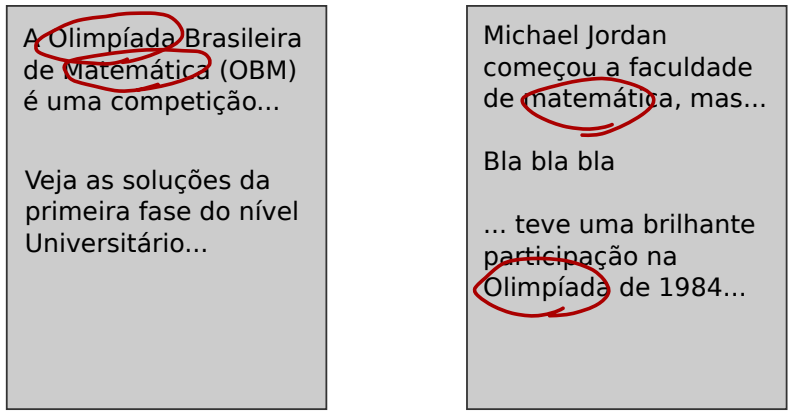
Um pouco sobre Ranking

- Em geral, as técnicas de ranking usadas pelos algoritmos de *ranking* anteriores eram focadas no **conteúdo das páginas**.
- Por exemplo, analisavam a proximidade das palavras procuradas, os títulos das páginas, etc.



Um pouco sobre Ranking

- Em geral, as técnicas de ranking usadas pelos algoritmos de *ranking* anteriores eram focadas no **conteúdo das páginas**.
- Por exemplo, analisavam a proximidade das palavras procuradas, os títulos das páginas, etc.



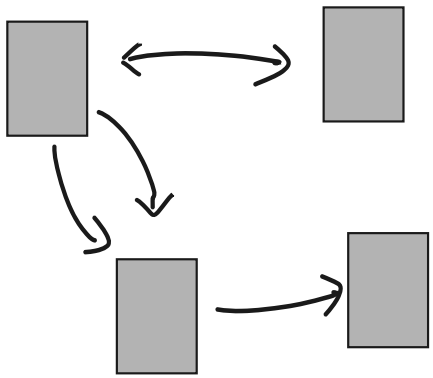
- A inovação de Page e Brin foi perceber que era possível usar a **estrutura da Web** para determinar a relevância das páginas.

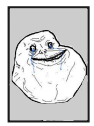
A Web como um grafo

Páginas na internet são conectadas por **hyperlinks** (ou apenas **links**). Podemos pensar nas páginas como nós de um grafo direcionado e nos links como as arestas.

A Web como um grafo

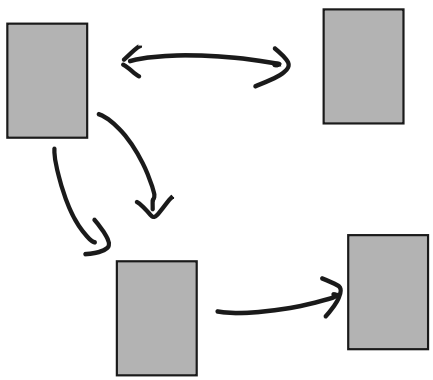
Páginas na internet são conectadas por **hyperlinks** (ou apenas **links**). Podemos pensar nas páginas como nós de um grafo direcionado e nos links como as arestas.

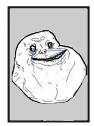




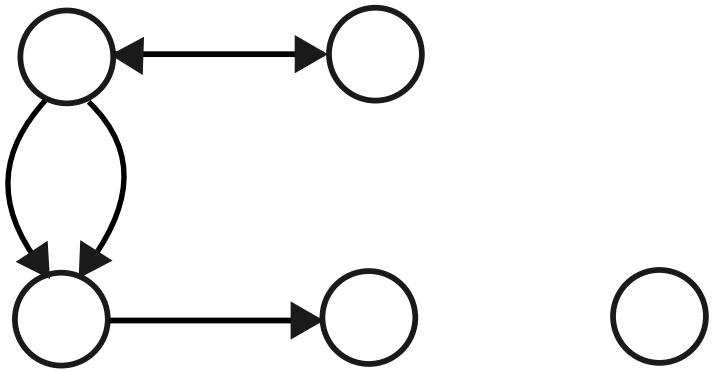
A Web como um grafo

Páginas na internet são conectadas por **hyperlinks** (ou apenas **links**). Podemos pensar nas páginas como nós de um grafo direcionado e nos links como as arestas.



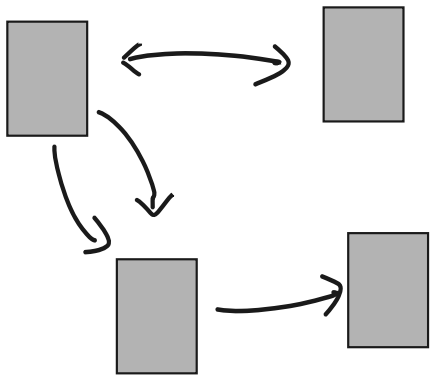


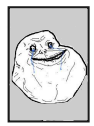
ou



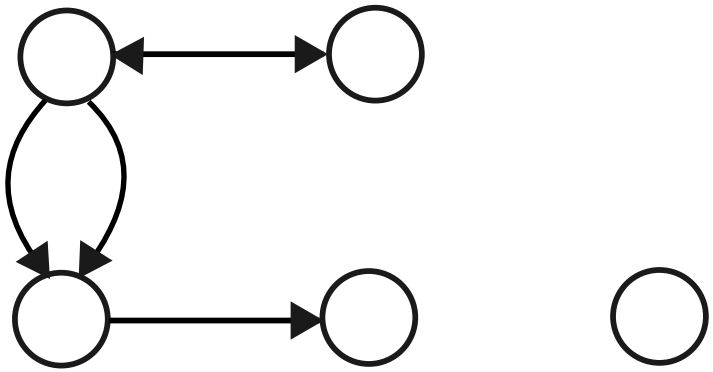
A Web como um grafo

Páginas na internet são conectadas por **hyperlinks** (ou apenas **links**). Podemos pensar nas páginas como nós de um grafo direcionado e nos links como as arestas.





ou



A ideia é simples: a **quantidade** de links que chegam e que saem de uma página, devem dizer alguma coisa sobre a relevância dela.

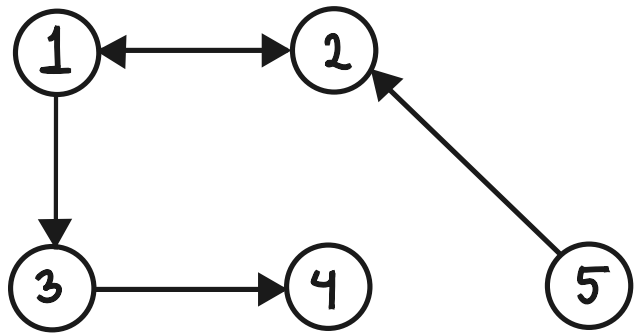
Notações e convenções

Notações e convenções

- Seja G um grafo direcionado, com nós 1,2,. . . ,n. Nosso objetivo é, para cada nó i, atribuir um número real x_i que traduza a importância ou relevância (no contexto da Web) do nó i.

Notações e convenções

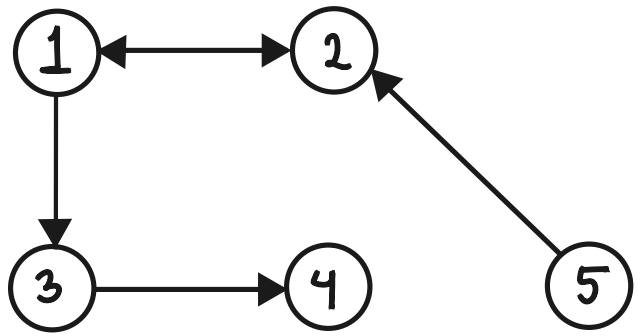
- Seja G um grafo direcionado, com nós 1,2,. . . ,n. Nosso objetivo é, para cada nó i, atribuir um número real x_i que traduza a importância ou relevância (no contexto da Web) do nó i.
- Quando j é um nó que possui uma aresta apontando para i, chamaremos de **link** para i e denotaremos $j \to i$. Por exemplo, no grafo



temos $1 \to 2, 3$, $2 \to 1$, $3 \to 4$, $5 \to 2$. (Obs.: Esta notação tem alguns problemas... mas vai ser útil assim mesmo).

Notações e convenções

- Seja G um grafo direcionado, com nós 1,2,. . . ,n. Nosso objetivo é, para cada nó i, atribuir um número real x_i que traduza a importância ou relevância (no contexto da Web) do nó i.
- Quando j é um nó que possui uma aresta apontando para i, chamaremos de **link** para i e denotaremos $j \to i$. Por exemplo, no grafo



temos $1 \to 2, 3$, $2 \to 1$, $3 \to 4$, $5 \to 2$. (Obs.: Esta notação tem alguns problemas... mas vai ser útil assim mesmo).

- O número de links saindo de j será denotado por l_j.

Definição de relevância: primeira tentativa

Definição de relevância: primeira tentativa

- Se uma página recebe muitos links, ela deve ser relevante!

Definição de relevância: primeira tentativa

- Se uma página recebe muitos links, ela deve ser relevante!Então poderíamos definir:

$$x_i = \sum_{j \to i} 1$$

Isto é, estamos apenas contando o número de links que apontam para i.

Definição de relevância: primeira tentativa

- Se uma página recebe muitos links, ela deve ser relevante!Então poderíamos definir:

$$x_i = \sum_{j \to i} 1$$

  Isto é, estamos apenas contando o número de links que apontam para i.
- **Problema**: Alguém poderia criar uma página com vários links para i, inflando artificialmente a sua relevância.

Definição de relevância: segunda tentativa

Definição de relevância: segunda tentativa

- É fácil resolver o problema anterior:

$$x_i = \sum_{j \to i} \frac{1}{l_j}$$

Definição de relevância: segunda tentativa

- É fácil resolver o problema anterior:

$$x_i = \sum_{j \to i} \frac{1}{l_j}$$

  Agora, se uma página aponta para i várias vezes, isto é levado em conta.

Definição de relevância: segunda tentativa

- É fácil resolver o problema anterior:

$$x_i = \sum_{j \to i} \frac{1}{l_j}$$

  Agora, se uma página aponta para i várias vezes, isto é levado em conta.
- **Problema**: Alguém poderia criar várias páginas "vazias", contendo um único link para i.

Definição de relevância: segunda tentativa

- É fácil resolver o problema anterior:

$$x_i = \sum_{j \to i} \frac{1}{l_j}$$

  Agora, se uma página aponta para i várias vezes, isto é levado em conta.
- **Problema**: Alguém poderia criar várias páginas "vazias", contendo um único link para i.
- Para entender como vamos resolver os dois problemas acima, vamos analisar um outro exemplo.

Um exemplo esclarecedor

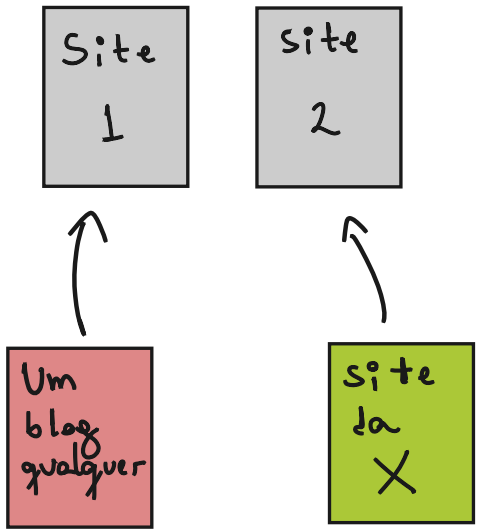
- Suponha que você quer comprar um produto eletrônico, fabricado pela empresa X, que não vende seus próprios produtos.

Um exemplo esclarecedor

- Suponha que você quer comprar um produto eletrônico, fabricado pela empresa X, que não vende seus próprios produtos.
- Você encontra dois sites que vendem o produto, que criativamente denominaremos Site 1 e Site 2.

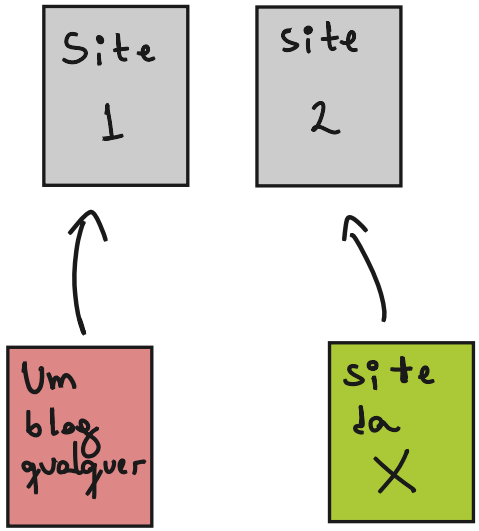
Um exemplo esclarecedor

- Suponha que você quer comprar um produto eletrônico, fabricado pela empresa X, que não vende seus próprios produtos.
- Você encontra dois sites que vendem o produto, que criativamente denominaremos Site 1 e Site 2.
- Queremos determinar qual destes dois sites é mais relevante, ou confiável, e cada um recebe apenas um link.



Um exemplo esclarecedor

- Suponha que você quer comprar um produto eletrônico, fabricado pela empresa X, que não vende seus próprios produtos.
- Você encontra dois sites que vendem o produto, que criativamente denominaremos Site 1 e Site 2.
- Queremos determinar qual destes dois sites é mais relevante, ou confiável, e cada um recebe apenas um link.



- O segundo link (verde) "vale mais" que o primeiro.

Definição de relevância: terceira tentativa

Definição de relevância: terceira tentativa

- A conclusão é que, para medir a relevância da página i, também devemos levar em conta a relevância das páginas j tais que $j \to i$.

Definição de relevância: terceira tentativa

- A conclusão é que, para medir a relevância da página i, também devemos levar em conta a relevância das páginas j tais que $j \to i$.
- Então poderíamos definir

$$x_i = \sum_{j \to i} \frac{1}{l_j} x_j$$

Definição de relevância: terceira tentativa

- A conclusão é que, para medir a relevância da página i, também devemos levar em conta a relevância das páginas j tais que $j \to i$.
- Então poderíamos definir

$$x_i = \sum_{j \to i} \frac{1}{l_j} x_j$$

- Parece promissor, mas ainda tem alguns **problemas**. Por exemplo, se uma página não é destino de nenhum link, ela deve ter relevância 0? Mais grave: o que acontece se uma página não emite links? Existe sempre uma solução?

Definição de relevância: terceira tentativa

- A conclusão é que, para medir a relevância da página i, também devemos levar em conta a relevância das páginas j tais que $j \to i$.
- Então poderíamos definir

$$x_i = \sum_{j \to i} \frac{1}{l_j} x_j$$

- Parece promissor, mas ainda tem alguns **problemas**. Por exemplo, se uma página não é destino de nenhum link, ela deve ter relevância 0? Mais grave: o que acontece se uma página não emite links? Existe sempre uma solução?
- Por ora, ~~vamos fazer como os físicos,~~ ignorar esses possíveis defeitos e tentar encontrar um método para achar a solução.

Formalizando melhor...

- O que conseguimos até agora foi um sistema linear $n \times n$:

$$\begin{cases} x_1 = \frac{m_{11}}{l_1} x_1 + \frac{m_{12}}{l_2} x_2 + \cdots + \frac{m_{1n}}{l_n} x_n \\ x_2 = \frac{m_{21}}{l_1} x_1 + \frac{m_{22}}{l_2} x_2 + \cdots + \frac{m_{2n}}{l_n} x_n \\ \vdots \\ x_n = \frac{m_{n1}}{l_1} x_1 + \frac{m_{n2}}{l_2} x_2 + \cdots + \frac{m_{nn}}{l_n} x_n \end{cases}$$

onde m_{ij} é o número de links $j \to i$ (pode ser 0).

Formalizando melhor...

- O que conseguimos até agora foi um sistema linear $n \times n$:

$$\begin{cases} x_1 = \frac{m_{11}}{l_1} x_1 + \frac{m_{12}}{l_2} x_2 + \cdots + \frac{m_{1n}}{l_n} x_n \\ x_2 = \frac{m_{21}}{l_1} x_1 + \frac{m_{22}}{l_2} x_2 + \cdots + \frac{m_{2n}}{l_n} x_n \\ \vdots \\ x_n = \frac{m_{n1}}{l_1} x_1 + \frac{m_{n2}}{l_2} x_2 + \cdots + \frac{m_{nn}}{l_n} x_n \end{cases}$$

  onde m_{ij} é o número de links $j \to i$ (pode ser 0).

- Usando um jargão mais matemático, se $A = (a_{ij})$ onde $a_{ij} = m_{ij}/l_j$, então podemos ver $A : \mathbb{R}^n \to \mathbb{R}^n$ como uma transformação linear e a **relevância x** $= [x_1, \ldots, x_n]$ é um autovetor do autovalor 1.

Formalizando melhor...

- O que conseguimos até agora foi um sistema linear $n \times n$:

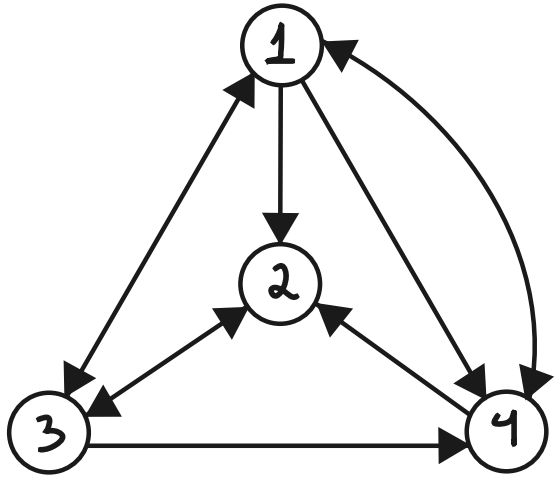
$$\begin{cases} x_1 = \frac{m_{11}}{l_1} x_1 + \frac{m_{12}}{l_2} x_2 + \cdots + \frac{m_{1n}}{l_n} x_n \\ x_2 = \frac{m_{21}}{l_1} x_1 + \frac{m_{22}}{l_2} x_2 + \cdots + \frac{m_{2n}}{l_n} x_n \\ \vdots \\ x_n = \frac{m_{n1}}{l_1} x_1 + \frac{m_{n2}}{l_2} x_2 + \cdots + \frac{m_{nn}}{l_n} x_n \end{cases}$$

  onde m_{ij} é o número de links $j \to i$ (pode ser 0).

- Usando um jargão mais matemático, se $A = (a_{ij})$ onde $a_{ij} = m_{ij}/l_j$, então podemos ver $A : \mathbb{R}^n \to \mathbb{R}^n$ como uma transformação linear e a **relevância** $\mathbf{x} = [x_1, \ldots, x_n]$ é um autovetor do autovalor 1. Ou, equivalentemente, um **ponto fixo** de A, i.e., $A\mathbf{x} = \mathbf{x}$.

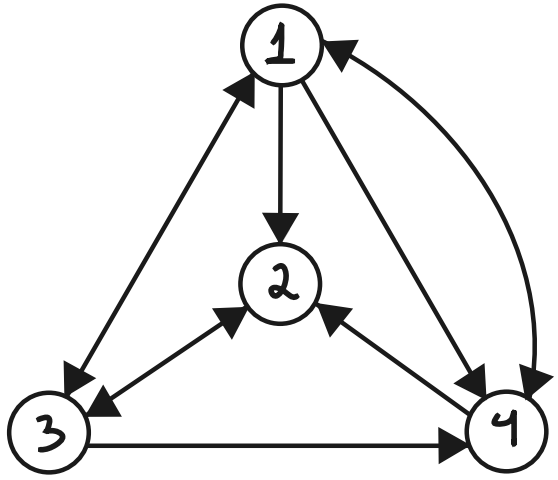
Exemplo

Exemplo



Neste caso, $l_1 = 4$, $l_2 = 1$, $l_3 = 3$ e $l_4 = 2$.

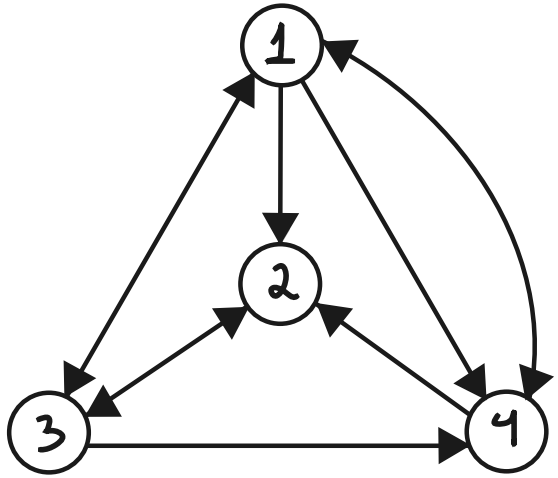
Exemplo



Neste caso, $l_1 = 4$, $l_2 = 1$, $l_3 = 3$ e $l_4 = 2$. O sistema linear fica:

$$\begin{cases} x_1 = \frac{0}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_2 = \frac{1}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_3 = \frac{1}{4}x_1 + \frac{1}{1}x_2 + \frac{0}{3}x_3 + \frac{0}{2}x_4 \\ x_4 = \frac{2}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{0}{2}x_4 \end{cases}$$

Exemplo



Neste caso, $l_1 = 4$, $l_2 = 1$, $l_3 = 3$ e $l_4 = 2$. O sistema linear fica:

$$\begin{cases} x_1 = \frac{0}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_2 = \frac{1}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_3 = \frac{1}{4}x_1 + \frac{1}{1}x_2 + \frac{0}{3}x_3 + \frac{0}{2}x_4 \\ x_4 = \frac{2}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{0}{2}x_4 \end{cases} \quad ou \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix}$$

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix}$$

- O vetor $[4, 5, 6, 4]$ indica, respectivamente, a relevância das páginas 1, 2, 3 e 4. Logo a página 3 é a mais relevante!

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix}$$

- O vetor $[4, 5, 6, 4]$ indica, respectivamente, a relevância das páginas 1, 2, 3 e 4. Logo a página 3 é a mais relevante!
- A solução não é única! Por exemplo, $[0.210526316, 0.263157895, 0.315789474, 0.210526316]$ também é solução.

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix}$$

- O vetor $[4, 5, 6, 4]$ indica, respectivamente, a relevância das páginas 1, 2, 3 e 4. Logo a página 3 é a mais relevante!
- A solução não é única! Por exemplo, $[0.210526316, 0.263157895, 0.315789474, 0.210526316]$ também é solução. Mas neste caso o espaço de soluções (autovetores do autovalor 1) tem dimensão 1, então tudo bem...

O que podemos dizer sobre esse sistema?

- É fácil achar uma solução:

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \\ 4 \end{bmatrix}$$

- O vetor $[4, 5, 6, 4]$ indica, respectivamente, a relevância das páginas 1, 2, 3 e 4. Logo a página 3 é a mais relevante!
- A solução não é única! Por exemplo, $[0.210526316, 0.263157895, 0.315789474, 0.210526316]$ também é solução. Mas neste caso o espaço de soluções (autovetores do autovalor 1) tem dimensão 1, então tudo bem...
- **Deveríamos estar nos perguntando**: sempre tem solução? e quanto à unicidade (ex: espaço de dimensão 2)?

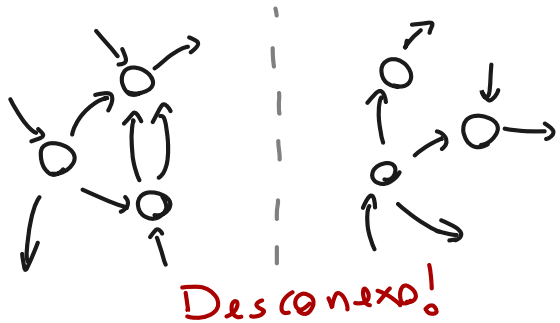
Bad news

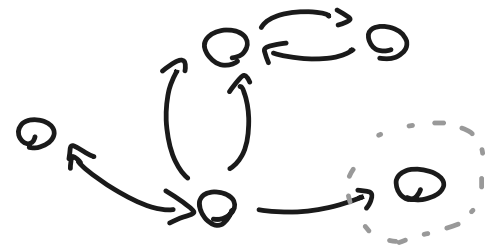
Bad news

- É fácil achar exemplos de grafos "estranhos" cujo espaço de soluções tenha dimensão maior ou igual a 2 ou sem solução para o sistema mencionado acima.

Bad news

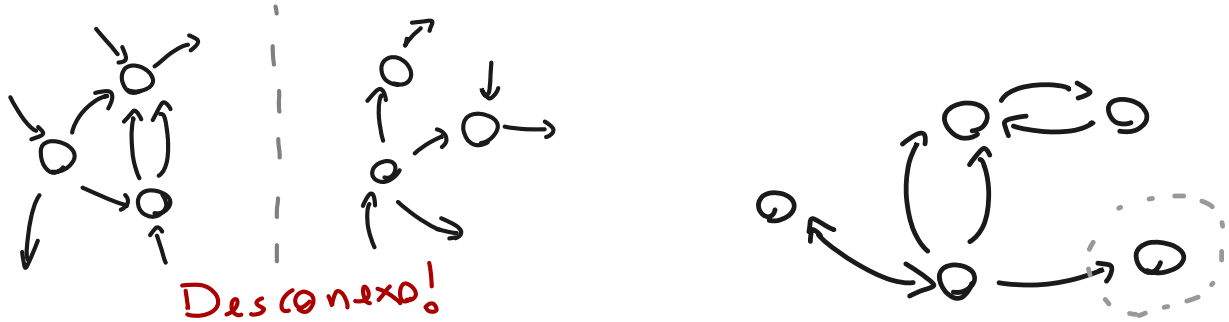
- É fácil achar exemplos de grafos "estranhos" cujo espaço de soluções tenha dimensão maior ou igual a 2 ou sem solução para o sistema mencionado acima.





Bad news

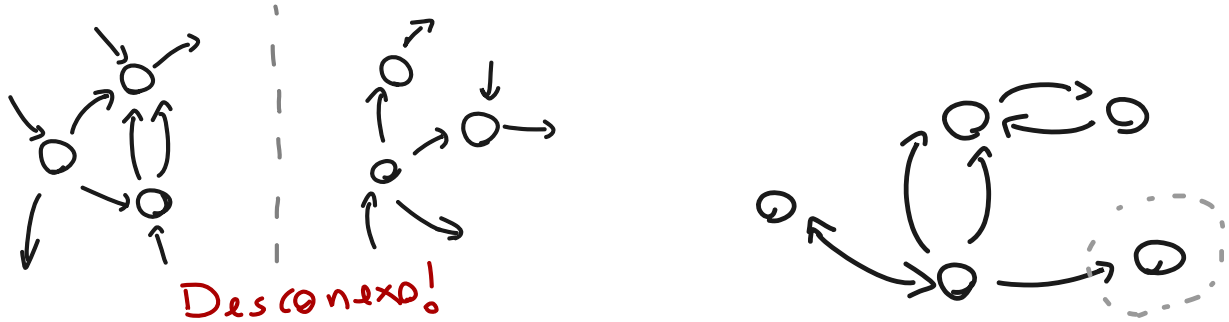
- É fácil achar exemplos de grafos "estranhos" cujo espaço de soluções tenha dimensão maior ou igual a 2 ou sem solução para o sistema mencionado acima.



- **Conclusão**: nosso modelo ainda não está muito bom, pois a priori deveria funcionar para um grafo qualquer (já que não temos controle sobre o grafo que representa a Web).

Bad news

- É fácil achar exemplos de grafos "estranhos" cujo espaço de soluções tenha dimensão maior ou igual a 2 ou sem solução para o sistema mencionado acima.



- **Conclusão**: nosso modelo ainda não está muito bom, pois a priori deveria funcionar para um grafo qualquer (já que não temos controle sobre o grafo que representa a Web).
- Mas nem tudo está perdido!

Interpretação probabilística: o internauta bêbado

Interpretação probabilística: o internauta bêbado

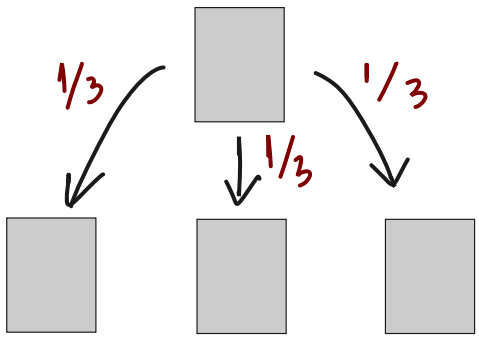
- Vamos entender melhor a matriz $A = (a_{ij})$, onde $a_{ij} = m_{ij}/l_j$.

Interpretação probabilística: o internauta bêbado

- Vamos entender melhor a matriz $A = (a_{ij})$, onde $a_{ij} = m_{ij}/l_j$.
- Podemos interpretar a_{ij} como sendo a probabilidade de, saindo do vértice j, chegar no vértice i!

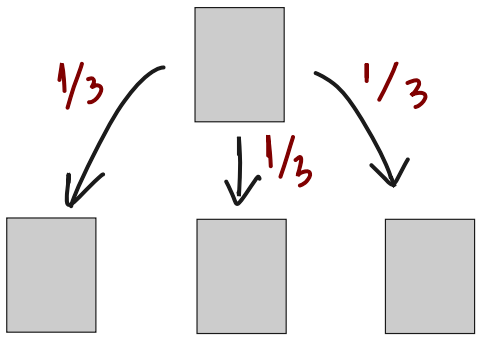
Interpretação probabilística: o internauta bêbado

- Vamos entender melhor a matriz $A = (a_{ij})$, onde $a_{ij} = m_{ij}/l_j$.
- Podemos interpretar a_{ij} como sendo a probabilidade de, saindo do vértice j, chegar no vértice i!
- No nosso caso concreto, isto quer dizer que estamos supondo que, se um internauta está na página j e clica em algum link aleatoriamente, ele possui probabilidade a_{ij} de chegar na página i.



Interpretação probabilística: o internauta bêbado

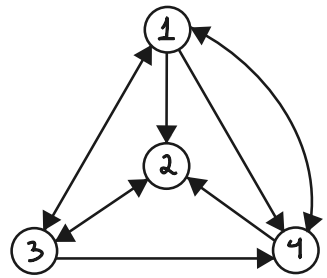
- Vamos entender melhor a matriz $A = (a_{ij})$, onde $a_{ij} = m_{ij}/l_j$.
- Podemos interpretar a_{ij} como sendo a probabilidade de, saindo do vértice j, chegar no vértice i!
- No nosso caso concreto, isto quer dizer que estamos supondo que, se um internauta está na página j e clica em algum link aleatoriamente, ele possui probabilidade a_{ij} de chegar na página i.



- Vejamos como a matriz (a_{ij}) deve ser utilizada na prática.

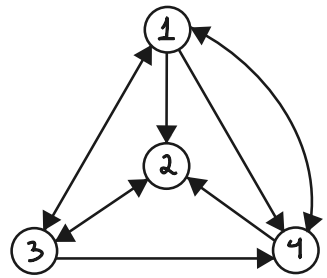
Voltando ao exemplo anterior

- Imagine alguém viajando em cima do grafo, onde em cada passo, o viajante vai de um nó para o outro, percorrendo uma aresta (no sentido indicado) que foi escolhida aleatóriamente.



Voltando ao exemplo anterior

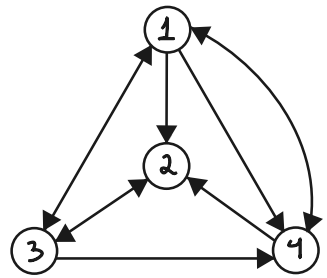
- Imagine alguém viajando em cima do grafo, onde em cada passo, o viajante vai de um nó para o outro, percorrendo uma aresta (no sentido indicado) que foi escolhida aleatóriamente.



Este procedimento é chamado de **passeio aleatório**.

Voltando ao exemplo anterior

- Imagine alguém viajando em cima do grafo, onde em cada passo, o viajante vai de um nó para o outro, percorrendo uma aresta (no sentido indicado) que foi escolhida aleatóriamente.

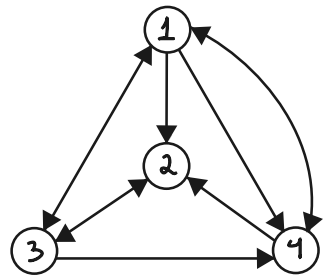


  Este procedimento é chamado de **passeio aleatório**.

- A matriz A nos diz "como devemos caminhar". Considere um vetor inicial $\mathbf{v}_0 = [1, 0, 0, 0]$, representando que o internauta começa na página 1.

Voltando ao exemplo anterior

- Imagine alguém viajando em cima do grafo, onde em cada passo, o viajante vai de um nó para o outro, percorrendo uma aresta (no sentido indicado) que foi escolhida aleatóriamente.



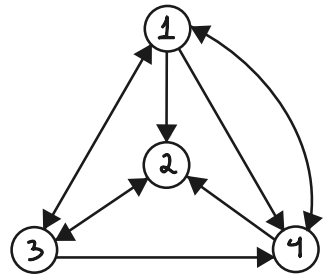
Este procedimento é chamado de **passeio aleatório**.

- A matriz A nos diz "como devemos caminhar". Considere um vetor inicial $\mathbf{v}_0 = [1, 0, 0, 0]$, representando que o internauta começa na página 1. Agora,

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix}$$

Voltando ao exemplo anterior

- Imagine alguém viajando em cima do grafo, onde em cada passo, o viajante vai de um nó para o outro, percorrendo uma aresta (no sentido indicado) que foi escolhida aleatóriamente.



  Este procedimento é chamado de **passeio aleatório**.

- A matriz A nos diz "como devemos caminhar". Considere um vetor inicial $\mathbf{v}_0 = [1, 0, 0, 0]$, representando que o internauta começa na página 1. Agora,

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix}$$

- A entrada i do vetor $\mathbf{v}_1 = [0, 1/4, 1/4, 1/2]$ é a probabilidade do internauta se encontrar na página i após ter clicado em algum link aleatoriamente da página 1.

Voltando ao exemplo anterior

- Repetindo o processo, temos

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/4 \\ 1/12 \end{bmatrix}$$

Agora, a entrada i do vetor $\mathbf{v}_2 = [1/3, 1/3, 1/4, 1/12]$ é a probabilidade do internauta se encontrar na página i após ter dado dois "cliques aleatórios". Etc...

Voltando ao exemplo anterior

- Repetindo o processo, temos

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/4 \\ 1/12 \end{bmatrix}$$

  Agora, a entrada i do vetor $\mathbf{v}_2 = [1/3, 1/3, 1/4, 1/12]$ é a probabilidade do internauta se encontrar na página i após ter dado dois "cliques aleatórios". Etc...

- Assim, a entrada i do vetor $\mathbf{v}_n$ é a probabilidade de encontrar o internauta (viajante) na página (nó) i, após n cliques (passos). Aqui, $\mathbf{v}_n = A\mathbf{v}_{n-1}$ e $\mathbf{v}_0 = [1, 0, 0, 0]$.

Voltando ao exemplo anterior

- Repetindo o processo, temos

$$\begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \\ 1/4 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/4 \\ 1/12 \end{bmatrix}$$

  Agora, a entrada i do vetor $\mathbf{v}_2 = [1/3, 1/3, 1/4, 1/12]$ é a probabilidade do internauta se encontrar na página i após ter dado dois "cliques aleatórios". Etc...

- Assim, a entrada i do vetor $\mathbf{v}_n$ é a probabilidade de encontrar o internauta (viajante) na página (nó) i, após n cliques (passos). Aqui, $\mathbf{v}_n = A\mathbf{v}_{n-1}$ e $\mathbf{v}_0 = [1, 0, 0, 0]$.
- As mesmas considerações valem para um grafo qualquer...

Simulando o passeio aleatório

Simulando o passeio aleatório

- Se simularmos um passeio aleatório (usando computador, por exemplo) com n passos, e contarmos o número de vezes N_i que o viajante passa pela página i, deveríamos ter $N_i/(n+1) \sim \mathbf{v}_n[i]$, onde $\mathbf{v}_n[i]$ é a i-ésima entrada de $\mathbf{v}_n$.

Simulando o passeio aleatório

- Se simularmos um passeio aleatório (usando computador, por exemplo) com n passos, e contarmos o número de vezes N_i que o viajante passa pela página i, deveríamos ter $N_i/(n+1) \sim \mathbf{v}_n[i]$, onde $\mathbf{v}_n[i]$ é a i-ésima entrada de $\mathbf{v}_n$.
- Por exemplo, para **100 passos**, obtive $N_1 = 23$, $N_2 = 25$, $N_3 = 32$ e $N_4 = 21$.

Simulando o passeio aleatório

- Se simularmos um passeio aleatório (usando computador, por exemplo) com n passos, e contarmos o número de vezes N_i que o viajante passa pela página i, deveríamos ter $N_i/(n+1) \sim \mathbf{v}_n[i]$, onde $\mathbf{v}_n[i]$ é a i-ésima entrada de $\mathbf{v}_n$.
- Por exemplo, para **100 passos**, obtive $N_1 = 23$, $N_2 = 25$, $N_3 = 32$ e $N_4 = 21$.
- Para **10000 passos**, $N_1 = 2101$, $N_2 = 2630$, $N_3 = 3161$ e $N_4 = 2109$. Ou melhor, calculando os valores $N_i/(n+1)$ aproximadamente:

$$0.210,\ 0.262,\ 0.316,\ 0.210$$

Simulando o passeio aleatório

- Se simularmos um passeio aleatório (usando computador, por exemplo) com n passos, e contarmos o número de vezes N_i que o viajante passa pela página i, deveríamos ter $N_i/(n+1) \sim \mathbf{v}_n[i]$, onde $\mathbf{v}_n[i]$ é a i-ésima entrada de $\mathbf{v}_n$.
- Por exemplo, para **100 passos**, obtive $N_1 = 23$, $N_2 = 25$, $N_3 = 32$ e $N_4 = 21$.
- Para **10000 passos**, $N_1 = 2101$, $N_2 = 2630$, $N_3 = 3161$ e $N_4 = 2109$. Ou melhor, calculando os valores $N_i/(n+1)$ aproximadamente:

$$0.210,\ 0.262,\ 0.316,\ 0.210$$

Ou seja, aproximadamente, um dos pontos fixos encontrados anteriormente!

$[\mathbf{0.210}526316, \mathbf{0.263}157895, \mathbf{0.315}789474, \mathbf{0.210}526316]$

Simulando o passeio aleatório

- Se simularmos um passeio aleatório (usando computador, por exemplo) com n passos, e contarmos o número de vezes N_i que o viajante passa pela página i, deveríamos ter $N_i/(n+1) \sim \mathbf{v}_n[i]$, onde $\mathbf{v}_n[i]$ é a i-ésima entrada de $\mathbf{v}_n$.
- Por exemplo, para **100 passos**, obtive $N_1 = 23$, $N_2 = 25$, $N_3 = 32$ e $N_4 = 21$.
- Para **10000 passos**, $N_1 = 2101$, $N_2 = 2630$, $N_3 = 3161$ e $N_4 = 2109$. Ou melhor, calculando os valores $N_i/(n+1)$ aproximadamente:

$$0.210,\ 0.262,\ 0.316,\ 0.210$$

Ou seja, aproximadamente, um dos pontos fixos encontrados anteriormente!

[**0.210**526316, **0.263**157895, **0.315**789474, **0.210**526316] Coincidência? Mágica?

O que podemos tirar disso?

O que podemos tirar disso?

- É possível provar, *sob certas condições no grafo e no vetor inicial*, que o que descobrimos acima vale em geral. Por ora, vamos pensar nisso apenas como uma **heurística**.

O que podemos tirar disso?

- É possível provar, *sob certas condições no grafo e no vetor inicial*, que o que descobrimos acima vale em geral. Por ora, vamos pensar nisso apenas como uma **heurística**.
- O que o internauta bêbado, ou melhor, o passeio aleatório tem a ver com a relevância das páginas?

O que podemos tirar disso?

- É possível provar, *sob certas condições no grafo e no vetor inicial*, que o que descobrimos acima vale em geral. Por ora, vamos pensar nisso apenas como uma **heurística**.
- O que o internauta bêbado, ou melhor, o passeio aleatório tem a ver com a relevância das páginas?
- Simples, os números N_i indicam uma medida de **popularidade** das páginas i (a página com o maior número de acessos, é a mais popular!).

O que podemos tirar disso?

- É possível provar, *sob certas condições no grafo e no vetor inicial*, que o que descobrimos acima vale em geral. Por ora, vamos pensar nisso apenas como uma **heurística**.
- O que o internauta bêbado, ou melhor, o passeio aleatório tem a ver com a relevância das páginas?
- Simples, os números N_i indicam uma medida de **popularidade** das páginas i (a página com o maior número de acessos, é a mais popular!).
- Agora temos duas concepções para a relevância de uma página, que intuimos serem a mesma coisa. No entanto, a concepção utilizando o passeio aleatório é **muito melhor** em vários sentidos.

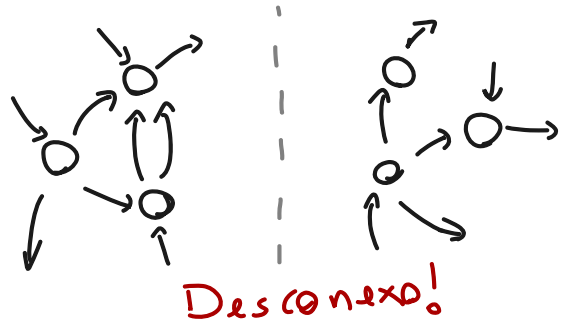
O que podemos tirar disso?

- É possível provar, *sob certas condições no grafo e no vetor inicial*, que o que descobrimos acima vale em geral. Por ora, vamos pensar nisso apenas como uma **heurística**.
- O que o internauta bêbado, ou melhor, o passeio aleatório tem a ver com a relevância das páginas?
- Simples, os números N_i indicam uma medida de **popularidade** das páginas i (a página com o maior número de acessos, é a mais popular!).
- Agora temos duas concepções para a relevância de uma página, que intuimos serem a mesma coisa. No entanto, a concepção utilizando o passeio aleatório é **muito melhor** em vários sentidos.
- Vejamos como esta nova interpretação explica os defeitos da nossa definição antiga de relevância e sugere como corrigí-la.

A explicação dos defeitos

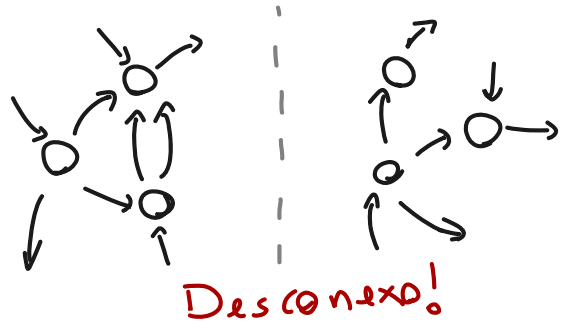
A explicação dos defeitos

No grafo desconexo, vemos que se o internauta começa em um dos pedaços, ele nunca pode passar para o outro.

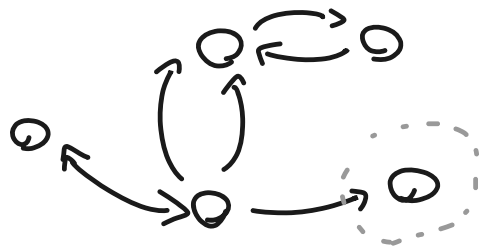


A explicação dos defeitos

No grafo desconexo, vemos que se o internauta começa em um dos pedaços, ele nunca pode passar para o outro.

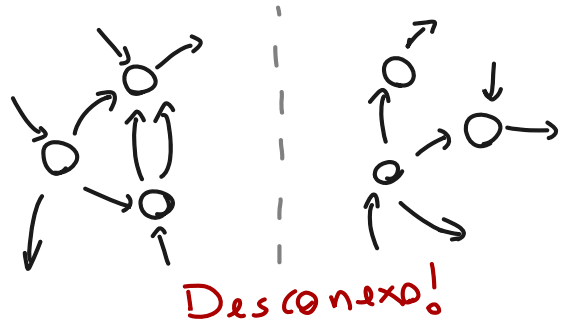


No segundo grafo, uma vez que o internauta chega na página sem links, ele fica "preso".

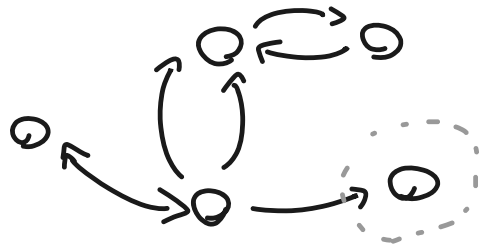


A explicação dos defeitos

No grafo desconexo, vemos que se o internauta começa em um dos pedaços, ele nunca pode passar para o outro.



No segundo grafo, uma vez que o internauta chega na página sem links, ele fica "preso".



É por isso que esses grafos produzem efeitos matemáticos estranhos na hora de encontrar a solução.

Uma sacada genial

Uma sacada genial

Uma sacada genial

- A ideia de Page e Brin para resolver os problemas acima e tornar o modelo um pouco mais realista foi introduzir o (que eu chamo de) **fator de entediamento**.

Uma sacada genial

- A ideia de Page e Brin para resolver os problemas acima e tornar o modelo um pouco mais realista foi introduzir o (que eu chamo de) **fator de entediamento**.
- Isto se baseia na seguinte ideia. Um internauta não fica passeando infinitamente pelos links: em algum momento ele fica entediado e pára de navegar, ou começa tudo de novo, em alguma outra página.

Uma sacada genial

- A ideia de Page e Brin para resolver os problemas acima e tornar o modelo um pouco mais realista foi introduzir o (que eu chamo de) **fator de entediamento**.
- Isto se baseia na seguinte ideia. Um internauta não fica passeando infinitamente pelos links: em algum momento ele fica entediado e pára de navegar, ou começa tudo de novo, em alguma outra página.
- Por exemplo, um estudante pode estar fazendo uma pesquisa para um trabalho escolar, mas alguma hora ele cansa e entra no Facebook ou no YouTube.

Uma sacada genial

- A ideia de Page e Brin para resolver os problemas acima e tornar o modelo um pouco mais realista foi introduzir o (que eu chamo de) **fator de entediamento**.
- Isto se baseia na seguinte ideia. Um internauta não fica passeando infinitamente pelos links: em algum momento ele fica entediado e pára de navegar, ou começa tudo de novo, em alguma outra página.
- Por exemplo, um estudante pode estar fazendo uma pesquisa para um trabalho escolar, mas alguma hora ele cansa e entra no Facebook ou no YouTube.
- Observe que isto resolve nossos problemas técnicos com os grafos, mas como introduzimos este fator no nosso modelo matemático?

Consertando o modelo

- Podemos, por exemplo, postular que sempre quando o internauta chega numa página i, ele possui uma probabilidade p de começar tudo de novo a partir de uma página escolhida aleatoreamente (e, consequentemente, uma probabilidade $1 - p$ de continuar seguindo os links). O Google utiliza $p = 0.15$.

Consertando o modelo

- Podemos, por exemplo, postular que sempre quando o internauta chega numa página i, ele possui uma probabilidade p de começar tudo de novo a partir de uma página escolhida aleatoreamente (e, consequentemente, uma probabilidade $1 - p$ de continuar seguindo os links). O Google utiliza $p = 0.15$.
- Assim, a aplicação que dita o caminho do viajante no passeio aleatório num grafo (de n vértices) dado é

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mapsto p \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} + (1-p) \begin{bmatrix} \frac{m_{11}}{l_1} & \frac{m_{12}}{l_2} & \cdots & \frac{m_{1n}}{l_n} \\ \frac{m_{21}}{l_1} & \frac{m_{22}}{l_2} & \cdots & \frac{m_{2n}}{l_n} \\ \vdots & & & \\ \frac{m_{n1}}{l_1} & \frac{m_{n2}}{l_2} & \cdots & \frac{m_{nn}}{l_n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Consertando o modelo

- Podemos, por exemplo, postular que sempre quando o internauta chega numa página i, ele possui uma probabilidade p de começar tudo de novo a partir de uma página escolhida aleatoreamente (e, consequentemente, uma probabilidade $1-p$ de continuar seguindo os links). O Google utiliza $p = 0.15$.
- Assim, a aplicação que dita o caminho do viajante no passeio aleatório num grafo (de n vértices) dado é

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mapsto p \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} + (1-p) \begin{bmatrix} \frac{m_{11}}{l_1} & \frac{m_{12}}{l_2} & \cdots & \frac{m_{1n}}{l_n} \\ \frac{m_{21}}{l_1} & \frac{m_{22}}{l_2} & \cdots & \frac{m_{2n}}{l_n} \\ \vdots & & & \\ \frac{m_{n1}}{l_1} & \frac{m_{n2}}{l_2} & \cdots & \frac{m_{nn}}{l_n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Que vamos denotar sinteticamente como

$$T : \mathbb{R}^n \to \mathbb{R}^n$$
$$\mathbf{y} \mapsto p\mathbf{e} + (1-p)A\mathbf{y}$$

Nova definição de relevância

- Podemos definir então, a **relevância**, ou a **popularidade**, ou o **PageRank**, da página i como sendo

$$x_i = p \cdot \frac{1}{n} + (1 - p) \cdot \sum_{j,\ j \to i} \frac{m_{ij}}{l_j} x_j$$

onde $p = 0.15$.

Nova definição de relevância

- Podemos definir então, a **relevância**, ou a **popularidade**, ou o **PageRank**, da página i como sendo

$$x_i = p \cdot \frac{1}{n} + (1 - p) \cdot \sum_{j,\ j \to i} \frac{m_{ij}}{l_j} x_j$$

  onde $p = 0.15$.
- Ou seja, um ponto fixo da aplicação acima.

Nova definição de relevância

- Podemos definir então, a **relevância**, ou a **popularidade**, ou o **PageRank**, da página i como sendo

$$x_i = p \cdot \frac{1}{n} + (1-p) \cdot \sum_{j,\ j \to i} \frac{m_{ij}}{l_j} x_j$$

  onde $p = 0.15$.
- Ou seja, um ponto fixo da aplicação acima.
- **Gambiarra**: se uma página j não possui nenhum link, postulamos no modelo que ela possui exatamente um link para cada outra página na Web. Isto força o internauta bêbado a recomeçar de outra página qualquer. Existem outras gambiarras possíveis, mas esta é a mais "justa".
- Esta é a fórmula usada no algoritmo do Google!

Nova definição de relevância

- Podemos definir então, a **relevância**, ou a **popularidade**, ou o **PageRank**, da página i como sendo

$$x_i = p \cdot \frac{1}{n} + (1-p) \cdot \sum_{j,\ j \to i} \frac{m_{ij}}{l_j} x_j$$

  onde $p = 0.15$.
- Ou seja, um ponto fixo da aplicação acima.
- **Gambiarra**: se uma página j não possui nenhum link, postulamos no modelo que ela possui exatamente um link para cada outra página na Web. Isto força o internauta bêbado a recomeçar de outra página qualquer. Existem outras gambiarras possíveis, mas esta é a mais "justa".
- Esta é a fórmula usada no algoritmo do Google!

Enfim, matemática!

- Vamos mostrar agora que a definição acima corresponde à noção do passeio aleatório. Além de ser a interpretação correta, isto nos fornecerá um meio eficiente para calcular a relevância das páginas.

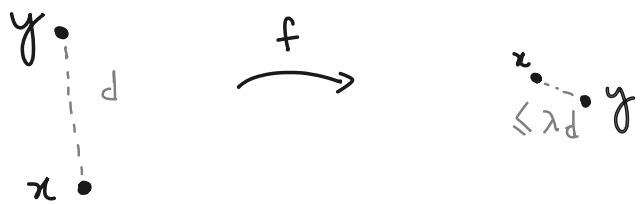
Enfim, matemática!

- Vamos mostrar agora que a definição acima corresponde à noção do passeio aleatório. Além de ser a interpretação correta, isto nos fornecerá um meio eficiente para calcular a relevância das páginas.
- Chegou a hora de usar matemática! :D

Definição

Uma função $f : \mathbb{R}^n \to \mathbb{R}^n$ *é uma* **contração** *se existe* $\lambda \in (0, 1)$ *tal que*

$$\|f(\mathbf{y}) - f(\mathbf{x})\| \leq \lambda \|\mathbf{y} - \mathbf{x}\|, \text{ para todos } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$



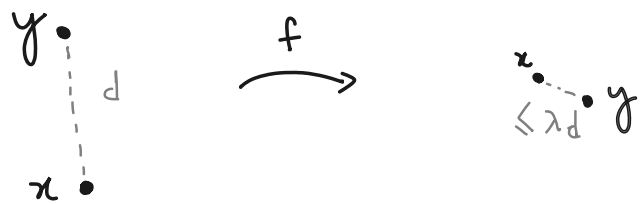
Enfim, matemática!

- Vamos mostrar agora que a definição acima corresponde à noção do passeio aleatório. Além de ser a interpretação correta, isto nos fornecerá um meio eficiente para calcular a relevância das páginas.
- Chegou a hora de usar matemática! :D

Definição

Uma função $f : \mathbb{R}^n \to \mathbb{R}^n$ *é uma* **contração** *se existe* $\lambda \in (0, 1)$ *tal que*

$$\|f(\mathbf{y}) - f(\mathbf{x})\| \leq \lambda \|\mathbf{y} - \mathbf{x}\|, \text{ para todos } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$



Exemplo bobo: $f : \mathbb{R}^2 \to \mathbb{R}^2$ tal que $f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} + (0, 1))$.

Teorema do Ponto Fixo de Banach

Teorema do Ponto Fixo de Banach

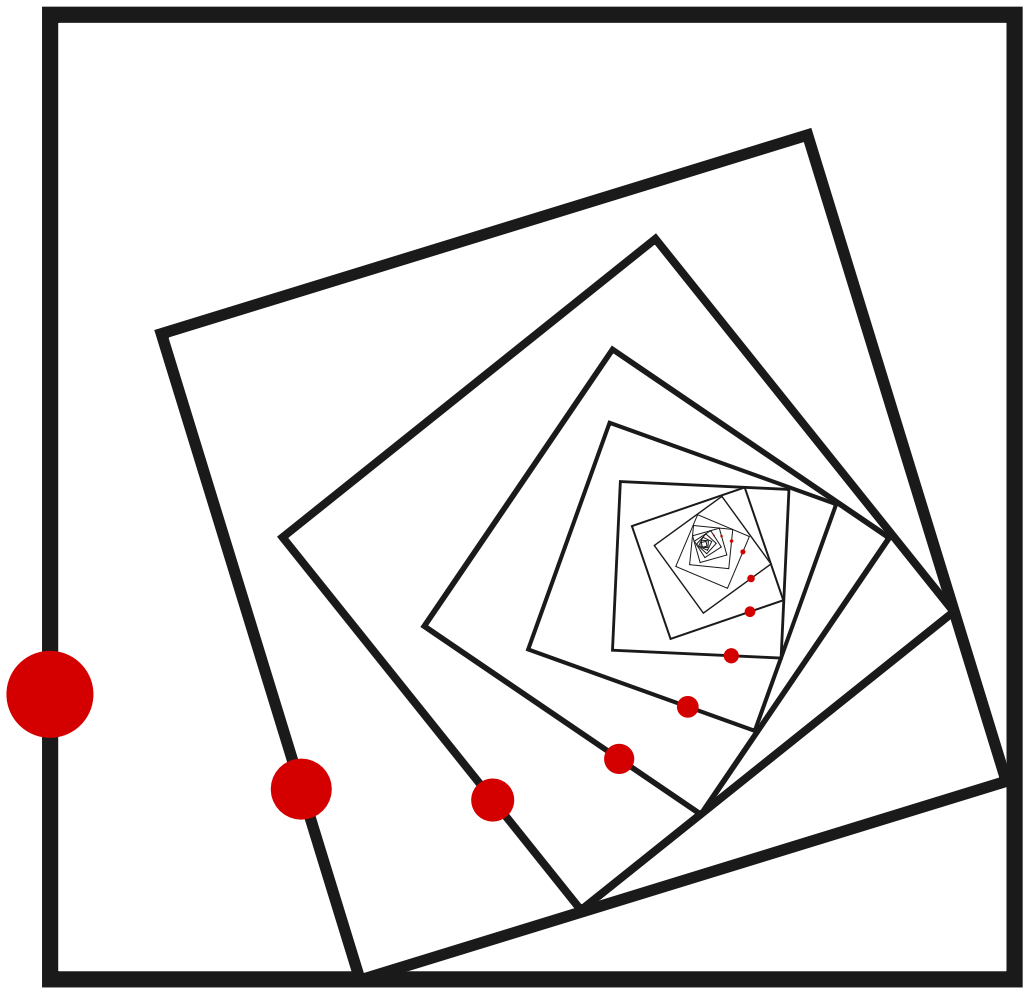
Teorema

Seja $f : \mathbb{R}^n \to \mathbb{R}^n$ *uma contração. Então existe um único* $\mathbf{p} \in \mathbb{R}^n$ *tal que* $f(\mathbf{p}) = \mathbf{p}$.

Teorema do Ponto Fixo de Banach

Teorema

Seja $f : \mathbb{R}^n \to \mathbb{R}^n$ *uma contração. Então existe um único* $\mathbf{p} \in \mathbb{R}^n$ *tal que* $f(\mathbf{p}) = \mathbf{p}$.



Demonstração do teorema

Demonstração.

- Seja $\lambda \in (0,1)$ tal que $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \lambda \|\mathbf{x} - \mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- **Unicidade**: Se $\mathbf{p} \neq \mathbf{q}$ são pontos fixos, então $\|\mathbf{p} - \mathbf{q}\| = \|f(\mathbf{p}) - f(\mathbf{q})\| \leq \lambda \|\mathbf{p} - \mathbf{q}\|$, o que implicaria $\lambda \geq 1$ pois $\|\mathbf{p} - \mathbf{q}\| > 0$. Absurdo.
- **Existência**: Seja $\mathbf{x}$ um ponto qualquer e defina a sequência $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{x}_n = f(\mathbf{x}_{n-1})$. Por indução, mostra-se que $\|\mathbf{x}_{m+1} - \mathbf{x}_m\| \leq \lambda^m \|\mathbf{x}_1 - \mathbf{x}_0\|$, $m \geq 1$. Então, dados $n, k \geq 1$, temos

$$\|\mathbf{x}_{n+k} - \mathbf{x}_n\| \leq \sum_{i=0}^{k-1} \|\mathbf{x}_{n+i+1} - \mathbf{x}_{n+i}\|$$

$$\leq \sum_{i=0}^{k-1} \lambda^{n+i} \|\mathbf{x}_1 - \mathbf{x}_0\| \leq \lambda^n \sum_{i=0}^{\infty} \lambda^i \|\mathbf{x}_1 - \mathbf{x}_0| = \frac{\lambda^n}{1-\lambda} \|\mathbf{x}_1 - \mathbf{x}_0\|$$

Como $0 < \lambda < 1$, isto mostra que $(\mathbf{x}_n)$ é uma sequência de Cauchy e, portanto, converge, digamos $\mathbf{x}_n \to \mathbf{p}$. Como f é contração, então f é contínua e segue que $f(\mathbf{p}) = f(\lim_n \mathbf{x}_n) = \lim_n f(\mathbf{x}_n) = \lim_n \mathbf{x}_{n+1} = \mathbf{p}$.

Alguns comentários

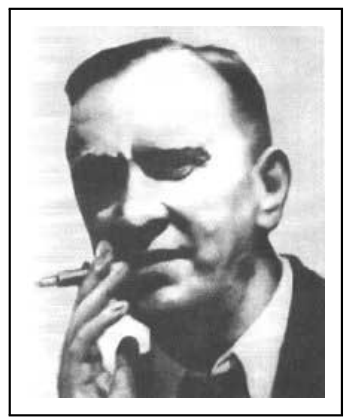
- Pela demonstração do teorema, podemos começar com **qualquer** ponto para obter o ponto fixo.

Alguns comentários

- Pela demonstração do teorema, podemos começar com **qualquer** ponto para obter o ponto fixo.
- O teorema vale em geral para espaços métricos completos e foi criado originalmente para resolver EDOs.

Alguns comentários

- Pela demonstração do teorema, podemos começar com **qualquer** ponto para obter o ponto fixo.
- O teorema vale em geral para espaços métricos completos e foi criado originalmente para resolver EDOs.
- Este é o Banach:



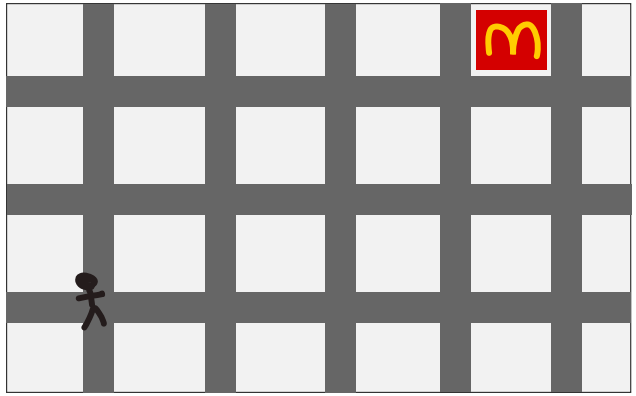
A norma da soma

A norma da soma

- Existem várias maneiras de medir distâncias!

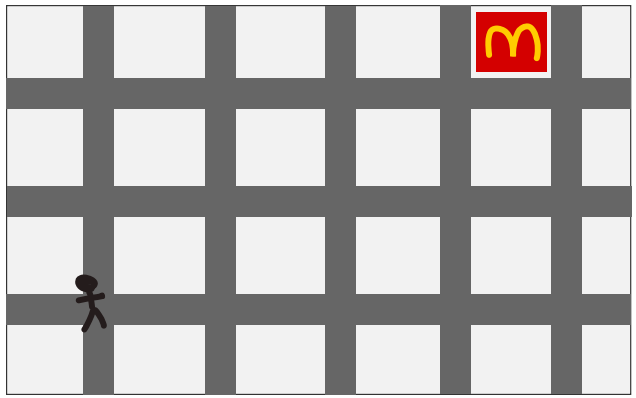
A norma da soma

- Existem várias maneiras de medir distâncias!
- Como você mediria a distância entre dois pontos numa cidade?

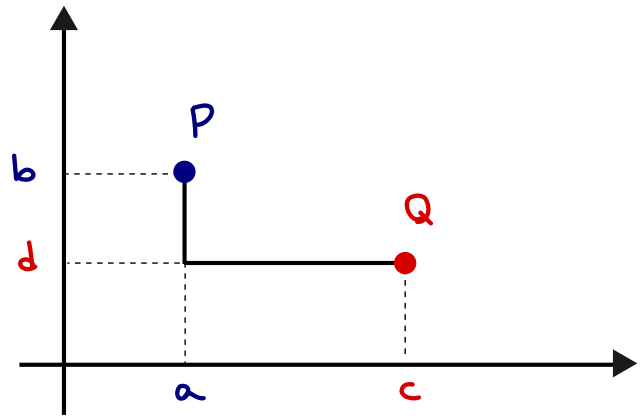


A norma da soma

- Existem várias maneiras de medir distâncias!
- Como você mediria a distância entre dois pontos numa cidade?



- Assim, no $\mathbb{R}^2$,



a distância entre $P = (a, b)$ e $Q = (c, d)$ é dada por $|a - c| + |b - d|$.

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.
3. $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, e $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$.

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.
3. $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, e $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$.

A norma $\|\cdot\|$ assim definida é chamada de **norma da soma**.

- Como é uma "bola" nessa norma?

A norma da soma

Definição

Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.
3. $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, e $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$.

A norma $\|\cdot\|$ assim definida é chamada de **norma da soma**.

- Como é uma "bola" nessa norma? Em $\mathbb{R}^2$, $\|(x, y)\| \leq 1 \iff |x| + |y| \leq 1$.

A norma da soma

Definição

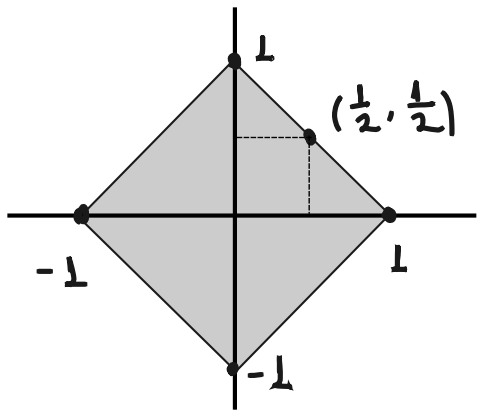
Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.
3. $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, e $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$.

A norma $\|\cdot\|$ assim definida é chamada de **norma da soma**.

- Como é uma "bola" nessa norma? Em $\mathbb{R}^2$, $\|(x, y)\| \leq 1 \iff |x| + |y| \leq 1$. Obtemos



A norma da soma

Definição

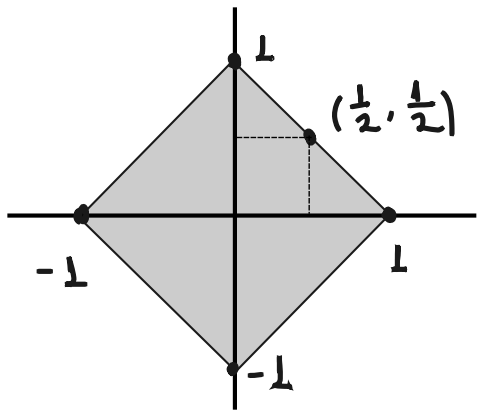
Se $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *definimos* $\|\mathbf{x}\| = |x_1| + \cdots + |x_n|$.

Note que $\|\cdot\|$ é uma **norma**, i.e., satisfaz:

1. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, $a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.
2. **(Desigualdade triangular)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}$.
3. $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, e $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$.

A norma $\|\cdot\|$ assim definida é chamada de **norma da soma**.

- Como é uma "bola" nessa norma? Em $\mathbb{R}^2$, $\|(x, y)\| \leq 1 \iff |x| + |y| \leq 1$. Obtemos



Aplicando ao nosso problema...

Aplicando ao nosso problema...

Mostremos agora que a aplicação $T : \mathbf{y} \mapsto p\mathbf{e} + (1-p)A\mathbf{y}$ é uma contração (considerando a norma da soma!).

Aplicando ao nosso problema...

Mostremos agora que a aplicação $T : \mathbf{y} \mapsto p\mathbf{e} + (1-p)A\mathbf{y}$ é uma contração (considerando a norma da soma!). Para isso, observe antes que:

- A matriz $A = (a_{ij})$, com $a_{ij} = m_{ij}/l_j$ é uma **matriz estocástica**, i.e.,
  1. Todas as suas entradas são positivas;
  2. A soma dos elementos de cada coluna é igual a 1: $\sum_{i=1}^{n} a_{ij} = \sum_{i=1}^{n} m_{ij}/l_j = l_j/l_j = 1$ para todo $1 \leq j \leq n$.
- $1 - p = 0.85$ é maior que 0 e menor que 1.

Aplicando ao nosso problema...

Mostremos agora que a aplicação $T : \mathbf{y} \mapsto p\mathbf{e} + (1-p)A\mathbf{y}$ é uma contração (considerando a norma da soma!). Para isso, observe antes que:

- A matriz $A = (a_{ij})$, com $a_{ij} = m_{ij}/l_j$ é uma **matriz estocástica**, i.e.,
  1. Todas as suas entradas são positivas;
  2. A soma dos elementos de cada coluna é igual a 1: $\sum_{i=1}^{n} a_{ij} = \sum_{i=1}^{n} m_{ij}/l_j = l_j/l_j = 1$ para todo $1 \leq j \leq n$.
- $1 - p = 0.85$ é maior que 0 e menor que 1.

Dados $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$,

$$\|T(\mathbf{y}) - T(\mathbf{z})\| = \|(1-p)A(\mathbf{y}-\mathbf{z})\| = (1-p)\sum_{j=1}^{n}\left(\sum_{i=1}^{n} a_{ij}|y_i - z_i|\right)$$

$$= (1-p)\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\right)|y_i - z_i|$$

$$= (1-p)\sum_{i=1}^{n}|y_i - z_i| = (1-p)\|\mathbf{y}-\mathbf{z}\|$$

Conclusão

Logo T é uma contração e, portanto,

- T possui um **único** ponto fixo **x**, e a relevância está bem definida.

Conclusão

Logo T é uma contração e, portanto,

- T possui um **único** ponto fixo **x**, e a relevância está bem definida.
- O Teorema mostra que podemos calcular **x** usando o passeio aleatório!

Conclusão

Logo T é uma contração e, portanto,

- T possui um **único** ponto fixo **x**, e a relevância está bem definida.
- O Teorema mostra que podemos calcular **x** usando o passeio aleatório!
- Se começarmos a aproximação com um ponto do tipo $\mathbf{x} = [a_0, \ldots, a_n]$, com $a_i \geq 0$ e $\sum_{i=1}^{n} a_i = 1$ isto quer dizer que nosso internauta bêbado começa na página i com probabilidade a_i.

Conclusão

Logo T é uma contração e, portanto,

- T possui um **único** ponto fixo $\mathbf{x}$, e a relevância está bem definida.
- O Teorema mostra que podemos calcular $\mathbf{x}$ usando o passeio aleatório!
- Se começarmos a aproximação com um ponto do tipo $\mathbf{x} = [a_0, \ldots, a_n]$, com $a_i \geq 0$ e $\sum_{i=1}^{n} a_i = 1$ isto quer dizer que nosso internauta bêbado começa na página i com probabilidade a_i. Calculando as iteradas de $\mathbf{x}$, i.e. $\mathbf{x}_1 = T(\mathbf{x})$, $\mathbf{x}_2 = T(T(\mathbf{x}))$, etc., cada $\mathbf{x}_i$ é tal que suas entradas são positivas e a soma das entradas é igual a 1.

Conclusão

Logo T é uma contração e, portanto,

- T possui um **único** ponto fixo $\mathbf{x}$, e a relevância está bem definida.
- O Teorema mostra que podemos calcular $\mathbf{x}$ usando o passeio aleatório!
- Se começarmos a aproximação com um ponto do tipo $\mathbf{x} = [a_0, \ldots, a_n]$, com $a_i \geq 0$ e $\sum_{i=1}^{n} a_i = 1$ isto quer dizer que nosso internauta bêbado começa na página i com probabilidade a_i. Calculando as iteradas de $\mathbf{x}$, i.e. $\mathbf{x}_1 = T(\mathbf{x})$, $\mathbf{x}_2 = T(T(\mathbf{x}))$, etc., cada $\mathbf{x}_i$ é tal que suas entradas são positivas e a soma das entradas é igual a 1. Um argumento simples com limite mostra que o mesmo vale para o ponto fixo de T. Ou seja, o ponto fixo também é uma distribuição de probabilidades no grafo inicial e a entrada i deve ser pensada como a frequência com que um internauta aleatório passa por i.

Resumindo

- O PageRank de uma página pode ser pensado como uma medida de popularidade. Ela é calculada usando um passeio aleatório no grafo direcionado que representa a Web com o truque adicional do fator de entediamento.

Resumindo

- O PageRank de uma página pode ser pensado como uma medida de popularidade. Ela é calculada usando um passeio aleatório no grafo direcionado que representa a Web com o truque adicional do fator de entediamento.
- Também pode ser interpretada como a relevância da página, levando em conta os links (recomendações) que apontam para ela, vindos de páginas relevantes.

Resumindo

- O PageRank de uma página pode ser pensado como uma medida de popularidade. Ela é calculada usando um passeio aleatório no grafo direcionado que representa a Web com o truque adicional do fator de entediamento.
- Também pode ser interpretada como a relevância da página, levando em conta os links (recomendações) que apontam para ela, vindos de páginas relevantes.
- O que garante que tudo funciona é o Teorema do Ponto Fixo de Banach, que diz que toda contração (em $\mathbb{R}^n$) tem um único ponto fixo, e calculamos este ponto fixo iterando a contração. Podemos começar com qualquer ponto.

Acabou?

Acabou?

Para um cientista da computação (ou um matemático aplicado), o problema acabou de começar!

Acabou?

Para um cientista da computação (ou um matemático aplicado), o problema acabou de começar!

- Criamos um modelo matemático para calcular a relevância da página i. Como implementá-lo **de maneira eficiente**?

Acabou?

Para um cientista da computação (ou um matemático aplicado), o problema acabou de começar!

- Criamos um modelo matemático para calcular a relevância da página i. Como implementá-lo **de maneira eficiente**?
- Estima-se que o tamanho do índice do Google está entre 40 e 45 bilhões de páginas. (`http://www.worldwidewebsize.com/`, 7 de Agosto de 2012)

Acabou?

Para um cientista da computação (ou um matemático aplicado), o problema acabou de começar!

- Criamos um modelo matemático para calcular a relevância da página i. Como implementá-lo **de maneira eficiente**?
- Estima-se que o tamanho do índice do Google está entre 40 e 45 bilhões de páginas. (`http://www.worldwidewebsize.com/`, 7 de Agosto de 2012)
- É impraticável armazenar toda a matriz A e realizar a multiplicão de matrizes usando o algoritmo usual, que possui complexidade n^2.

Acabou?

Para um cientista da computação (ou um matemático aplicado), o problema acabou de começar!

- Criamos um modelo matemático para calcular a relevância da página i. Como implementá-lo **de maneira eficiente**?
- Estima-se que o tamanho do índice do Google está entre 40 e 45 bilhões de páginas. (`http://www.worldwidewebsize.com/`, 7 de Agosto de 2012)
- É impraticável armazenar toda a matriz A e realizar a multiplicão de matrizes usando o algoritmo usual, que possui complexidade n^2.
- O que fazer? A matriz A é uma **matriz esparsa** (i.e, a maioria de seus elementos são 0), existem algoritmos específicos para lidar com esse tipo de matriz.

Problemas computacionais

- Por exemplo, no nosso caso, economizaríamos tempo se usássemos a fórmula direto (sem uso de matrizes): seja L_j o "conjunto" das páginas para as quais j aponta, incluindo repetições (em particular, $|L_j| = l_j$).

Problemas computacionais

- Por exemplo, no nosso caso, economizaríamos tempo se usássemos a fórmula direto (sem uso de matrizes): seja L_j o "conjunto" das páginas para as quais j aponta, incluindo repetições (em particular, $|L_j| = l_j$).

  Entrada: x
  Saída: $y = Tx$
  —
  Inicialize $y_i = p/n$, *para todo* $i = 1, \ldots, n$.
  Para j *de* 1 *até* n *faça*
  Para i *em* L_j *faça*
  $y_i = y_i + \frac{1-p}{l_j} x_j$

  Agora o algoritmo tem ordem n.

Problemas computacionais

- Por exemplo, no nosso caso, economizaríamos tempo se usássemos a fórmula direto (sem uso de matrizes): seja L_j o "conjunto" das páginas para as quais j aponta, incluindo repetições (em particular, $|L_j| = l_j$).

  Entrada: x
  Saída: $y = Tx$
  —
  Inicialize $y_i = p/n$, *para todo* $i = 1, \ldots, n$.
  Para j *de* 1 *até* n *faça*
  Para i *em* L_j *faça*
  $y_i = y_i + \frac{1-p}{l_j} x_j$

  Agora o algoritmo tem ordem n.

- Existem também um outros problemas pois calculamos o ponto fixo de maneira aproximada. Quando parar de iterar o processo? E os erros "acumulados"?

Problemas computacionais

- Se torturarmos o Teorema do Ponto Fixo de Banach, eles nos diz também qual é a velocidade da convergência e como estimar o erro da aproximação!

Problemas computacionais

- Se torturarmos o Teorema do Ponto Fixo de Banach, eles nos diz também qual é a velocidade da convergência e como estimar o erro da aproximação!

Proposição

Se $f : \mathbb{R}^n \to \mathbb{R}^n$ é uma contração com coeficiente $\lambda \in (0,1)$, $x = x_0$ é um ponto qualquer e p é o ponto fixo de f, então $\|x_n - p\| \leq \lambda^n \|x_0 - p\|$ e, mais precisamente,

$$\|x_n - p\| \leq \frac{\lambda}{1-\lambda} \|x_n - x_{n-1}\|$$

onde $x_m = f(x_{m-1})$, $m \geq 1$.

Problemas computacionais

- Se torturarmos o Teorema do Ponto Fixo de Banach, eles nos diz também qual é a velocidade da convergência e como estimar o erro da aproximação!

Proposição

Se $f : \mathbb{R}^n \to \mathbb{R}^n$ é uma contração com coeficiente $\lambda \in (0, 1)$, $x = x_0$ é um ponto qualquer e p é o ponto fixo de f, então $\|x_n - p\| \leq \lambda^n \|x_0 - p\|$ e, mais precisamente,

$$\|x_n - p\| \leq \frac{\lambda}{1 - \lambda} \|x_n - x_{n-1}\|$$

onde $x_m = f(x_{m-1})$, $m \geq 1$.

- Outra coisa interessante é: o grafo que representa a Web é constantemente atualizado, porém, poucas páginas são mudadas em uma atualização. Isto sugere que a relevância das páginas deve mudar pouco em uma atualização.

Problemas computacionais

- Se torturarmos o Teorema do Ponto Fixo de Banach, eles nos diz também qual é a velocidade da convergência e como estimar o erro da aproximação!

Proposição

Se $f : \mathbb{R}^n \to \mathbb{R}^n$ é uma contração com coeficiente $\lambda \in (0, 1)$, $x = x_0$ é um ponto qualquer e p é o ponto fixo de f, então $\|x_n - p\| \leq \lambda^n \|x_0 - p\|$ e, mais precisamente,

$$\|x_n - p\| \leq \frac{\lambda}{1 - \lambda} \|x_n - x_{n-1}\|$$

onde $x_m = f(x_{m-1})$, $m \geq 1$.

- Outra coisa interessante é: o grafo que representa a Web é constantemente atualizado, porém, poucas páginas são mudadas em uma atualização. Isto sugere que a relevância das páginas deve mudar pouco em uma atualização. Então podemos aproveitar a relevância antiga para calcular a nova relevância em muito menos passos! (Pelo teorema, não importa o ponto inicial).

Curiosidades

Curiosidades

- A fórmula do PageRank está errada no artigo original.

Curiosidades

- A fórmula do PageRank está errada no artigo original.
- O buscador da Google utiliza ~ 250 técnicas diferentes de ranking combinadas.

Curiosidades

- A fórmula do PageRank está errada no artigo original.
- O buscador da Google utiliza $\sim$ 250 técnicas diferentes de ranking combinadas.
- No artigo original, também é dito que é quase impossível enganar o Google (i.e., inflar artificialmente a relevância da sua página). Isto é notoriamente **falso**. Por exemplo, em 2011 o New York Times descobriu uma empresa que sempre aparecia no topo de diversas pesquisas e suspeitou de que ela se utilizasse de más práticas, o que acabou sendo comprovado após uma perícia da Google. Veja o artigo "*The Dirty Little Secrets of Search*"
  `http://www.nytimes.com/2011/02/13/business/13search.html?r=1`

Curiosidades

- A fórmula do PageRank está errada no artigo original.
- O buscador da Google utiliza $\sim$ 250 técnicas diferentes de ranking combinadas.
- No artigo original, também é dito que é quase impossível enganar o Google (i.e., inflar artificialmente a relevância da sua página). Isto é notoriamente **falso**. Por exemplo, em 2011 o New York Times descobriu uma empresa que sempre aparecia no topo de diversas pesquisas e suspeitou de que ela se utilizasse de más práticas, o que acabou sendo comprovado após uma perícia da Google. Veja o artigo "*The Dirty Little Secrets of Search*"
  `http://www.nytimes.com/2011/02/13/business/13search.html?r=1`
- A Google tenta ficar atenta a este tipo de prática e pune os sites quando necessário.

Curiosidades

- A fórmula do PageRank está errada no artigo original.
- O buscador da Google utiliza $\sim$ 250 técnicas diferentes de ranking combinadas.
- No artigo original, também é dito que é quase impossível enganar o Google (i.e., inflar artificialmente a relevância da sua página). Isto é notoriamente **falso**. Por exemplo, em 2011 o New York Times descobriu uma empresa que sempre aparecia no topo de diversas pesquisas e suspeitou de que ela se utilizasse de más práticas, o que acabou sendo comprovado após uma perícia da Google. Veja o artigo "*The Dirty Little Secrets of Search*"
  `http://www.nytimes.com/2011/02/13/business/13search.html?r=1`
- A Google tenta ficar atenta a este tipo de prática e pune os sites quando necessário.
- No site da Google existe um manual de boas práticas ensinando como melhorar o seu PageRank honestamente.

Curiosidades

- A fórmula do PageRank está errada no artigo original.
- O buscador da Google utiliza $\sim$ 250 técnicas diferentes de ranking combinadas.
- No artigo original, também é dito que é quase impossível enganar o Google (i.e., inflar artificialmente a relevância da sua página). Isto é notoriamente **falso**. Por exemplo, em 2011 o New York Times descobriu uma empresa que sempre aparecia no topo de diversas pesquisas e suspeitou de que ela se utilizasse de más práticas, o que acabou sendo comprovado após uma perícia da Google. Veja o artigo "*The Dirty Little Secrets of Search*"
  `http://www.nytimes.com/2011/02/13/business/13search.html?r=1`
- A Google tenta ficar atenta a este tipo de prática e pune os sites quando necessário.
- No site da Google existe um manual de boas práticas ensinando como melhorar o seu PageRank honestamente.
- Eu usei o Google 47 vezes para fazer essa apresentação.

Referências

Referências

http://www.google.com

Referências (agora é sério...)

- *Nine Algorithms That Changed the Future: The Ingenious Ideas That Drive Today's Computers*, JOHN MACCORMICK. Livro muito interessante, escrito para leigos, explicando em linguagem simples as ideias por trás de alguns dos algoritmos mais utilizados hoje em dia.
- *Comment fonctionne Google ?*, MICHAEL EISERMANN, `http://www.igt.uni-stuttgart.de/eiserm/enseignement/google.pdf`. Artigo expositório de 15 páginas no qual eu baseei esta apresentação. Muito bem escrito, mas em francês . . .
- *PageRank*, artigo da Wikipedia, `http://en.wikipedia.org/wiki/Pagerank`. Além de explicar mais ou menos o algoritmo, fala um pouco da sua história. Além disso, possui desenhos muito melhores que os meus e outros diversos links interessantes.
- *The anatomy of a large-scale hypertextual Web search engine*, S. BRIN & L. PAGE, Computer Networks and ISDN Systems 30: 107–117, `http://infolab.stanford.edu/pub/papers/google.pdf`. Artigo original explicando o funcionamento do Google.