PageRank

Pagerank

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites

Pagerank

La idea es simular un navegante de la web aleatorio, que hace clicks y va cambiando de páginas

La idea es calcular en qué páginas es más probable que termine

El algoritmo incluye un *damping factor* **d**, que corresponde a la probabilidad que deje de hacer links y salte a una página aleatoriamente

Pagerank

- **Entrada**: grafo
- **Salida**: distribución de probabilidad relativa de que un usuario haciendo clicks random termine en una cierta página web
- Loops (links de una página a sí misma) son ignorados. Múltiples links de una página a otra son considerados como un mismo arco

Pagerank

- Inicialmente, todas las páginas tienen la misma probabilidad

- En cada iteración, cada página le entrega parte de su probabilidad a todas aquellas a las que tiene un link

- En cada iteración, la probabilidad de un nuevo salto se atenúa según el factor *d*

Pagerank

Algoritmo

En la iteración 0 definimos el PageRank para cada nodo como:

$$PR_0(n_i) = 1/N$$

Y para las siguientes iteraciones como:

$$PR_t(n_i) = \frac{1-d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR_{t-1}(n_j)}{Out(n_j)}$$

Pagerank

Algoritmo

$$PR_t(n_i) = \frac{1-d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR_{t-1}(n_j)}{Out(n_j)}$$

PageRank del nodo i en la iteración t

Número de nodos en el grafo

Damping factor, usualmente 0.85

Nodos que tienen una arista entrando al nodo i

Número de aristas saliendo del nodo j

Pagerank

Algoritmo

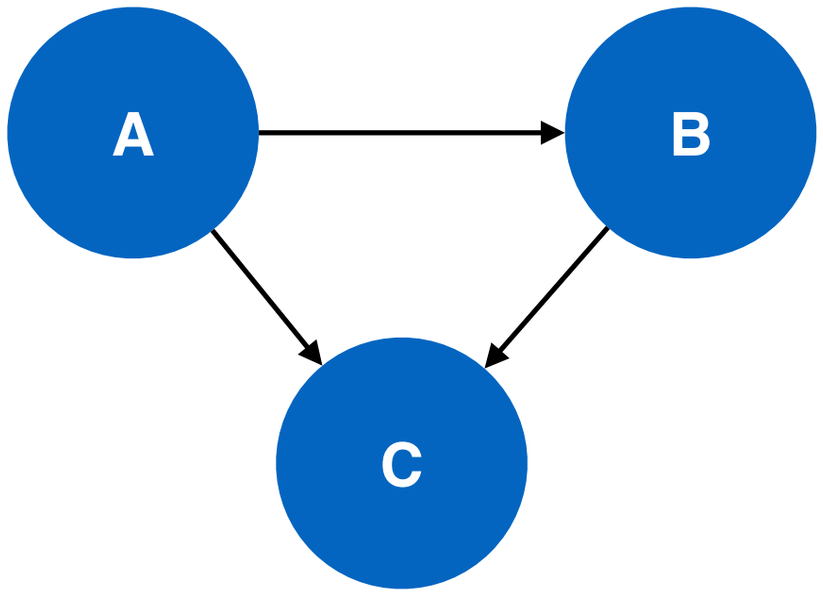
En cada iteración calculamos el siguiente valor:

$$\sqrt{(PR_t(nodo_1) - PR_{t-1}(nodo_1))^2 + \ldots + (PR_t(nodo_N) - PR_{t-1}(nodo_N))^2}$$

Hasta que tome un valor "suficientemente pequeño"

Pagerank

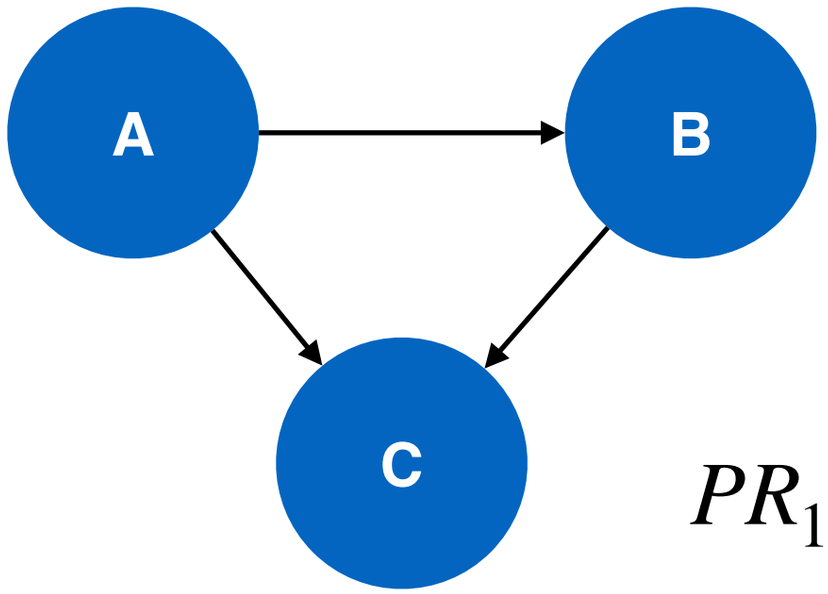
Ejemplo



$$PR_0(\text{A}) = PR_0(\text{B}) = PR_0(\text{C}) = \frac{1}{3}$$

Pagerank

Ejemplo



$$PR_0(\text{A}) = PR_0(\text{B}) = PR_0(\text{C}) = \frac{1}{3}$$

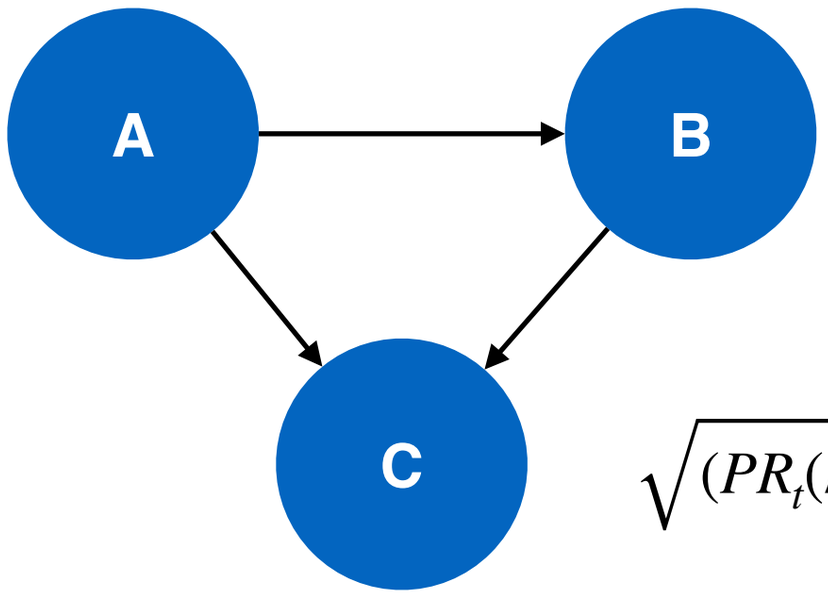
$$PR_1(\text{A}) = \frac{1 - 0.85}{3} + 0.85 \cdot 0$$

$$PR_1(\text{B}) = \frac{1 - 0.85}{3} + 0.85 \cdot (\frac{PR_0(\text{A})}{2})$$

$$PR_1(\text{C}) = \frac{1 - 0.85}{3} + 0.85 \cdot (\frac{PR_0(\text{A})}{2} + \frac{PR_0(\text{B})}{1})$$

Pagerank

Ejemplo



Iteramos hasta que el valor de:

$$\sqrt{(PR_t(\mathrm{A}) - PR_{t-1}(\mathrm{A}))^2 + (PR_t(\mathrm{B}) - PR_{t-1}(\mathrm{B}))^2 + (PR_t(\mathrm{C}) - PR_{t-1}(\mathrm{C}))^2}$$

Sea menor a un número muy bajo (por ejemplo, 0.0001)

¿Qué pasa con la web?

¿Qué pasa con la web?

- Si d es muy grande, necesitamos muchas iteraciones para converger

¿Qué pasa con la web?

- Si d es muy grande, necesitamos muchas iteraciones para converger
- Si d es muy chico, todo converge a 1

¿Qué pasa con la web?

- Si d es muy grande, necesitamos muchas iteraciones para converger
- Si d es muy chico, todo converge a 1
- En el paper original, el grafo de la web converge aceptablemente en 52 iteraciones con d = 0.85