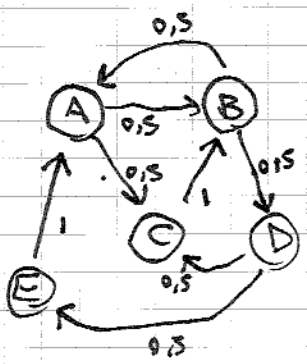
PageRank

Busca ranking nodos en un grafo en base a su importancia.



Matriz adyacencia →

	A	B	C	D	E
A	0	1/2	0	0	1
B	1/2	0	1	0	0
C	1/2	0	0	1/2	0
D	0	1/2	0	0	0
F	0	0	0	1/2	0

→ M

Los pesos en los aristas de un nodo son 1 sobre la cantidad de aristas que salen del nodo. La matriz se forma con los pesos. Sea m_{ij} la pos (i,j) de la matriz entonces m_{ij} es el peso de la arista que va de j a i.

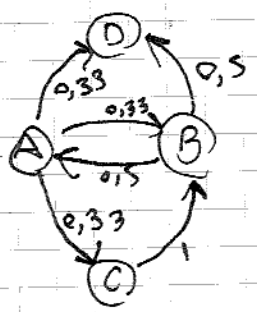
Todos los nodos empiezan con un $PR = \frac{1}{V}$ donde V es la cant. de vértices. Cada nodo reparte en partes iguales su PR a todos los nodos que linkea. El nuevo PR de cada nodo se calcula como la suma de todas las porciones de PR recibidas por otros nodos. Esto se repite hasta que el PR converge en el valor real.

Sea X_n el PR del nodo X en la iteración N:

$$\begin{bmatrix} A_n \\ B_n \\ C_n \\ D_n \\ E_n \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1 \\ 1/2 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \times \begin{bmatrix} A_{n-1} \\ B_{n-1} \\ C_{n-1} \\ D_{n-1} \\ E_{n-1} \end{bmatrix}$$

con $X_0 = \frac{1}{5}$.

Dead ends



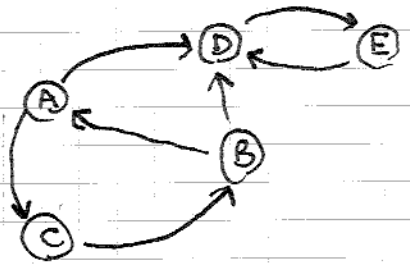
$$\longrightarrow \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

Un dead end es un nodo que no tiene aristas de salida (columna de ceros en M). A la larga este nodo acumulará todo el PR. Para resolver esto distribuimos equitativamente todo su PR con todos los otros nodos (incluyendose).

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

PageRank. 2

Spider Traps y Teletransportación.



Podemos ver que en este caso D. y E acumulan todo el PR entre los dos porque hacen un ciclo cerrado entre ellos. Esto se llama Spider Trap y para resolverlo utilizamos el concepto de Teletransportación.

En vez de siempre mandar el PR a los vecinos lo hacemos con una probabilidad β. En el caso que no salga (probabilidad de $1-\beta$) enviamos todo el PR a un nodo al azar. Esto se puede simular cambiando la fórmula de PR, en este caso la fórmula quedaría como:

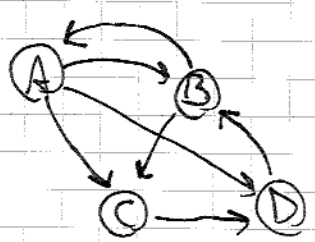
$$PR = \beta \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \cdot PR + (1-\beta) \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix}$$

Topic Rank.

Buscamos rankear las páginas que corresponden a cierto tema.

Dada una consulta debemos determinar su "tema". Buscamos Teletransportarnos únicamente a las páginas que están clasificadas bajo dicho tema. Para esto asignamos $(1-\beta)\cdot\frac{1}{|C|}$ a dichas páginas siendo $|C|$ la cantidad total de páginas dentro del tema.

Un ejemplo:



Si A y C tienen el mismo tema que nuestra consulta entonces:

$$PR = \beta \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \end{bmatrix} + (1-\beta) \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \\ 0 \end{bmatrix}$$

Trust Rank

Mismo procedimiento que Topic rank pero solo Teletransportamos a páginas que confiamos.

PageRank 3

¿Cómo si una página es confiable?

- Selección manual (intervención humana)
- Páginas con muchos clicks luego de búsquedas
- Páginas de dominios .edu, .gov, .mil, etc.

La masa de spam nos da una idea de cuanto del page rank es spam y podemos eliminar páginas que superen cierto umbral.

$$\text{Masa de spam} = \frac{PR - PT}{PR}$$

PR: page rank
PT: Trust rank.

Sim Rank

Buscamos utilizar page rank para determinar cuales son las páginas más semejantes a una dada. Para esto calculamos PR pero modificamos la teletransportación para solo teletransportarnos a nuestra página de interés. De esta forma, el PR nos dará un ranking de semejanza a nuestra página.

Podemos aproximar el sim rank usando Montecarlo. Se realizan "n" random walks a partir de cada nodo y se suma 1 a cada nodo visitado. El nodo inicial siempre es el que nos interesa encontrar similares.

Visual rank

Queremos poder generar un ranking de imágenes. Intuitivamente, sabemos que la imagen más parecida a todas las demás debería ser la más relevante y así sucesivamente. Con todas las imágenes recuperadas se construye un grafo completo. La probabilidad de cada link depende de la semejanza entre las imágenes. Las imágenes con mayor page rank en este grafo son las que primero mostramos.

Para calcular la semejanza usamos "scale invariant feature transform" (SIFT).

Text rank.

Igual que en visual rank pero usando semejanzas de texto. Tiene aplicaciones bastante interesantes:

- Generación de resúmenes automáticos
- Extracción de keywords.