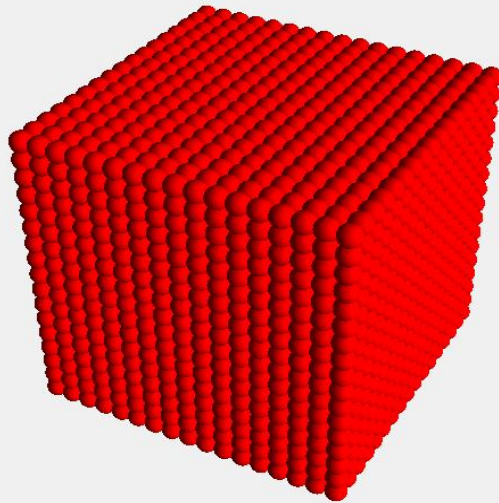


Page Rank



Contenido

1. Page Rank	2
1.1. Introducción	2
1.2. Matriz de conectividad	2
1.3. Calculo del PageRank	3
1.4. Algo de definiciones	5
1.5. Ejercicios	6

1

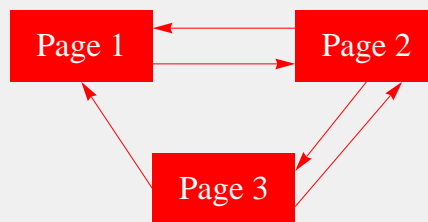
Page Rank

1.1. Introducción

En esta ocasión nos dedicamos a estudiar el problema de “Page Rank”, que es un algoritmo que lista las páginas web en orden de importancia. Obviamente pueden existir diferentes criterios para dar una medida de importancia a una página web, sin embargo aquí explicaremos el que se usa en el algoritmo usado por Google.

1.2. Matriz de conectividad

Para comenzar nuestro estudio pensemos en un baby-Internet que consiste solo de tres páginas web y que están conectadas como muestra la figura, es decir, en la página 1 existen una liga para que se acceda a la página 2, de manera similar en la página 2 existe una liga para poder acceder a la página 1. Ocurre lo mismo para las páginas 2 y 3. Para las páginas 1 y 3, solo existe un botón en 3 para ir a 1.



Matriz de conectividad: es una matriz C_{ij} donde $c_{ij} = 0$ si no existe una liga de la página j a la página i y $c_{ij} = \frac{1}{k}$ donde k es el número total de ligas que parten de la página j siempre y cuando exista una liga a la página i .

Para nuestro caso tenemos una matriz de conectividad 3×3 .

$$M_t = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}$$

Ahora cómo podemos medir la importancia de nuestras páginas. Para iniciar el estudio supongamos que las tres tienen la misma importancia definiendo al vector PageRank inicial, que para nuestro caso es:

$$PR_0 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$$

1.3. Calculo del PageRank

Para la página 1 Veamos que ocurre con el pagerank de la página 1,

De la página 2 Recibe 1/2 del pagerank de la 2, debido a que la página 2 tiene 2 salidas.

De la página 3 Recibe también 1/2 de la página 3 ya que tiene dos salidas.

Por lo tanto la página 1 tiene un pagerank de $(1/2)(1/3)$ de 3 y $(1/2)(1/3)$ de 2, después de la primera transición, es decir un total de $(1/2)(1/3) + (1/2)(1/3) = 1/3$.

Para la página 2 Veamos que ocurre con el pagerank de la página 2,

De la página 1 Recibe 1 de esta página, debido a que esta página solo tiene una salida.

De la página 3 Recibe 1/2 de esta página, debido a que esta página tiene 2 salidas.

Por lo tanto la página 2 tiene un pagerank de $(1/2)(1/3)$ de 1 y $(1)(1/3)$ de 3, después de la primera transición. Es decir un total de $(1)(1/3) + (1/2)(1/3) = 1/2$.

Para la página 3 Veamos que ocurre con el pagerank de la página 3,

De la página 1 No recibe nada de esta página.

De la página 2 Recibe $(1/2)$, debido a que la página 2 tiene dos ligas de salida.

Por lo tanto la página 3 tiene un pagerank de $(0)(1/3)$ de 1 y $(1/2)(1/3)$ de 3, después de la primera transición. Es decir un total de $(1/2)(1/3) = 1/6$.

En resumen el nuevo vector Pagerank es:

$$PR_1 = \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix}$$

De manera matricial tenemos que:

$$M_t PR_0 = PR_1$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,5000 \\ 0,1666 \end{pmatrix}$$

Sin embargo no hay razón para pensar que PR_2 sea mejor medida que PR_1 . Entonces ampliamos el número de transiciones, y veamos que sucede.

$$M_t PR_1 = PR_2$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 5/12 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4166 \\ 0,2500 \end{pmatrix}$$

$$M_t PR_2 = PR_3$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 5/12 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 11/24 \\ 5/24 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4583 \\ 0,2083 \end{pmatrix}$$

$$M_t PR_3 = PR_4$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 11/24 \\ 5/24 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 7/16 \\ 11/48 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4375 \\ 0,2291 \end{pmatrix}$$

$$M_t PR_4 = PR_5$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 7/16 \\ 11/48 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 43/96 \\ 7/32 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4479 \\ 0,2187 \end{pmatrix}$$

$$M_t PR_5 = PR_6$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 43/96 \\ 7/32 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 85/192 \\ 43/192 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4427 \\ 0,2239 \end{pmatrix}$$

$$M_t PR_6 = PR_7$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 85/192 \\ 43/192 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 57/128 \\ 85/384 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4453 \\ 0,2213 \end{pmatrix}$$

$$M_t PR_7 = PR_8$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 57/128 \\ 85/384 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 341/768 \\ 57/256 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4440 \\ 0,2226 \end{pmatrix}$$

$$M_t PR_8 = PR_9$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 341/768 \\ 57/256 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 683/1536 \\ 341/1536 \end{pmatrix} = \begin{pmatrix} 0,3333 \\ 0,4446 \\ 0,2220 \end{pmatrix}$$

Lo anterior nos permite inferir que esta sucesión converge al vector:

$$PR = \begin{pmatrix} 0,333 \\ 0,444 \\ 0,222 \end{pmatrix}$$

Además observamos que:

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 0,333 \\ 0,444 \\ 0,222 \end{pmatrix} = \begin{pmatrix} 0,333 \\ 0,444 \\ 0,222 \end{pmatrix}$$

Es decir PR es un vector propio de la matriz de conectividad M_t asociado al valor propio 1.

$$(M_t)(PR) = (1)(PR)$$

1.4. Algo de definiciones



Ahora estamos en condiciones de formalizar algunos conceptos anteriores.

Definición 1 Una red Internet es una gráfica dirigida, donde en cada nodo sale al menos una dirección y no existen direcciones que salen y entran al mismo nodo.

Definición 2 Sea una red Internet con n vértices y sea k_j el número de aristas que tienen como origen el vértice P_j . La matriz de conectividad de la red Internet es la matriz $n \times n$ $A = (a_{ij})$, donde:

$$a_{ij} = \begin{cases} \frac{1}{k_j} & \text{si existe una arista de } P_j \text{ a } P_i \\ 0 & \text{en otro caso} \end{cases}$$

Teorema 1 Si A es una matriz de conectividad de una red Internet, entonces:

1. $a_{ii} = 0$.
2. Las entradas de cada columna suma 1.
3. El número $\lambda = 1$ es un valor propio de A .

Definición 3 Si A es la matriz de conectividad de una red Internet, sea $X = (x_1, x_2, \dots, x_n)$ un valor propio de A asociado con el valor propio $\lambda = 1$, y sea $s = x_1 + x_2 + \dots + x_n$. Si $s \neq 0$, entonces un simple PageRank de la red Internet es el vector $(1/s)X$.

1.5. Ejercicios

Encontrar la matriz de conectividad y un simple PageRank de las siguientes redes Internet.

