DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP 1: PageRank

September 6, 2022 Métodos Numéricos

Grupo 18

Integrante	LU	Correo electrónico
Vekselman, Natán	338/21	natanvek11@gmail.com
Arienti, Federico	316/21	fa.arienti@gmail.com
Barcos, Juan Cruz	463/20	juancruzbarcos@hotmail.com

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (++54 +11) 4576-3300

http://www.exactas.uba.ar

RESUMEN

El Ranking de Page, *PageRank* [1], es un método propuesto por Sergey Brin y Larry Page —co-fundadores de Google—, para establecer la importancia de una página web dentro del internet, o dentro de un subconjunto de las páginas que lo componen. Holísticamente, el ranking calcula el puntaje de cada página como la fracción de tiempo, al largo plazo, que un navegante permanecerá en ella [2].

Desde una perspectiva algorítmica, PageRank busca resolver un sistema lineal $\mathbf{A}x = x$, donde $\mathbf{A}$ es una matriz estocástica en columnas [2] y cada una de sus posiciones a_{ij} representa la probabilidad de que un usuario situado en la página j decida navegar a la página i.

Este trabajo propone una implementación eficiente del ranking a través del uso de distintas representaciones de matríz (acorde a su 'ralidad'), y el empleo de iteradores específicos, para reducir el costo espacial y temporal de la eliminación gaussiana, método utilizado para la resolución del sistema.

Se buscará dar una presentación teórica y una evaluación cuantitativa y cualitativa de los resultados de tanto el método propuesto, como de PageRank en si.

Palabras clave: *Ranking de Page, Eliminación Gaussiana, Matrices ralas*

Contenidos

1. Introducción Teórica

1.1. **Aridad.** Consideremos primero el dominio y la imágen de PageRank.

Dominio: 1. un conjunto de páginas web interconectadas a través de hipervínculos. Podemos considerar este conjunto como un grafo direccionado, donde los nodos son los sitios y los ejes, los links. 2. un parámetro de entrada $p \in (0,\ 1)$, que representa la probabilidad que un usuario decida navegar aleatoriamente a otra página en el grafo. Se puede interpretar como el parámetro de un variable aleatoria de Bernoulli.

Imágen: un vector $x \in [0,\ 1]^n$, donde x_i representa el Ranking de Page para la i-ésima página del conjunto de entrada, donde x satisface que $x_i \geq 0\ \forall i : 0\ ...\ n$ y $\sum_{i=1}^{n} x_i = 1$.

Tenemos entonces:

$$PageRank :\ G_n \times (0,\ 1) \longrightarrow [0,\ 1]^n \qquad \forall n \in \mathbb{N} \tag{1}$$

donde G_n refiere al conjunto de conjuntos de páginas web, interconectadas a través de hipervínculos, con cardinalidad n.

1.2. **El sistema.** PageRank propone resolver un sistema de ecuaciones para encontrar la relevancia de cada página i $(i : 1\ ...\ n)$ en $g \in G_n$:

$$x_i := \sum_{j=1}^{n} x_j \cdot Pr(j \longrightarrow i) \tag{2}$$

donde $Pr(j \longrightarrow i)$ es la probabilidad que un usuario situado en la página j decida ir a la página i. Se define de la siguente manera:

$$Pr(j \longrightarrow i) := \begin{cases} (1-p) \cdot \frac{1}{n} + p \cdot \frac{I_{ij}}{c_j} & \text{si} \quad c_j \neq 0 \\ \frac{1}{n} & \text{si no} \end{cases} \tag{3}$$

con $I_{ij} = 1$ si y sólo si existe un hipervínculo de j a i, con $j \neq i$ —y nulo en caso contrario—, y $c_j = \sum_{i=1}^{n} I_{ij}$, la cantidad de links salientes de j. La restricción $j \neq i$ será para evitar que se consideren auto-referencias en el ranking.

Notemos que $Pr(j \longrightarrow i)$ se puede interpretar de la siguente manera: un navegante situado en la página j decidirá con probabilidad p acceder a uno de los links del sitio y con probabilidad $1 - p$ saltar a otra página del conjunto. En ambos casos, deberá luego decidir uniformemente sobre el total disponible, y terminará eligiendo a i con una probabilidad de $\frac{I_{ij}}{c_j}$ ó $\frac{1}{n}$, respectivamente, acorde a la primer decisión. Si no hay links en la página, siempre saltará de manera uniforme a otra página del conjunto, y eligirá a i con probabilidad $\frac{1}{n}$.

x_i, por su parte, también recibe una interpretación particular: es la probabilidad que para algún momento $k > K$, el navegante se encuentre situado en la página i. Para un K lo suficientemente grande, esta probabilidad es única [3].

A este modelo se lo conoce como el *modelo del navegante aleatorio*. El mismo asume lo siguiente: un link de la página j a la página i es evidencia de la importancia de la página i. Específicamente: la cantidad de relevancia que le confiere la página j a la página i es proporcional a la relevancia de j e inversamente proporcional al número de páginas a las que apunta j [3]. Esto se puede ver en que x_i es la suma de toda otra importancia x_j ponderada por $Pr(j \longrightarrow i)$, que incluye en su definición una división sobre la cantidad de links salientes c_j.

1.3. **Representación matricial.** Dado que estamos trabajando con un sistema lineal, será de utilidad considerar la matriz asociada $\boldsymbol{A}$ y resolver, equivalentemente, $\mathbf{A}x = x$. Definimos entonces:

$$a_{ij} := Pr(j \longrightarrow i) \tag{4}$$

y proponemos que[1]:

$$\mathbf{A} = p\mathbf{W}\mathbf{D} + ez^t \tag{5}$$

donde, $\forall i,\ j:\ 1 \ldots n$, se satisface que:

$$e_i = 1$$

$$z_j = \begin{cases} (1-p)/n & \text{si} \quad c_j \neq 0 \\ 1/n & \text{si no} \end{cases}$$

$$w_{ij} = \begin{cases} 1 & \text{si} \quad i \neq j \ \wedge \ j \xrightarrow{l} i \\ 0 & \text{si no} \end{cases}$$

$$d_{ij} = \begin{cases} 1/c_j & \text{si} \quad i = j \ \wedge \ c_j \neq 0 \\ 0 & \text{si no} \end{cases}$$

La notación $j \xrightarrow{l} i$ representa que existe un link de la página j a la página i, y las filas y columnas de $\mathbf{W}$, denominada *matriz de conectividad*, representan —indexadas por posición— las páginas de g.

A partir de esta última equivalencia podemos ver que:

[1]Una demostración de esta equivalencia se encuentra en 5.1.

$$\begin{aligned} \mathbf{A}x &= x \\ (p\mathbf{WD} + ez^t)x &= x \\ p\mathbf{WD}x + ez^t x &= x \\ x - p\mathbf{WD}x &= ez^t x \\ (\mathbf{I} - p\mathbf{WD})x &= \gamma e \end{aligned}$$

donde $\gamma = z^t x$.

Si asumimos $\gamma = 1$, entonces el sistema a resolver será:

$$(\mathbf{I} - p\mathbf{WD})x = e \tag{6}$$

El sistema definido en (6) permite la aplicación de la eliminación gaussiana sin permutación.[2] Por lo que utilizaremos éste método para su resolución. Como no necesariamente $\sum_{i=1}^{n} x_i = 1$, normalizaremos el resultado alcanzado para satisfacer los requerimientos de la imágen.

[2]Una demostración de este enunciado se encuentra en 5.2.

2. Desarrollo

3. Resultados y Discusión

4. Conclusiones

5. Apéndice

5.1. $A = pWD + ez^t$.

demostración. Recordemos que:

$$e_i = 1$$

$$z_j = \begin{cases} (1-p)/n & \text{si} \quad c_j \neq 0 \\ 1/n & \text{si no} \end{cases}$$

$$w_{ij} = \begin{cases} 1 & \text{si} \quad i \neq j \ \wedge \ j \xrightarrow{l} i \\ 0 & \text{si no} \end{cases}$$

$$d_{ij} = \begin{cases} 1/c_j & \text{si} \quad i = j \ \wedge \ c_j \neq 0 \\ 0 & \text{si no} \end{cases}$$

A partir de estas definiciones, vemos que, como **D** es diagonal, el producto a derecha **WD** escala cada columna w_j por el factor d_{jj}, tal que:

$$(\mathbf{WD})_{ij} = \begin{cases} w_{ij}/c_j & \text{si} \quad c_j \neq 0 \\ 0 & \text{si no} \end{cases}$$

Como p es un escalar, sigue entonces que:

$$(p\mathbf{WD})_{ij} = \begin{cases} p \cdot w_{ij}/c_j & \text{si} \quad c_j \neq 0 \\ 0 & \text{si no} \end{cases}$$

Además, $e \in \mathbb{R}^{n\times 1} \ \wedge \ z^t \in \mathbb{R}^{1\times n} \implies ez^t \in \mathbb{R}^{n\times n}$, y:

$$(ez^t)_{ij} := \sum_{k=1}^{1} e_{ik} \cdot z^t_{kj} = e_i \cdot z^t_j = 1 \cdot z^t_j = z_j$$

Por lo que:

$$(p\mathbf{WD} + ez^t)_{ij} = \begin{cases} p \cdot w_{ij}/c_j + z_j & \text{si} \quad c_j \neq 0 \\ z_j & \text{si no} \end{cases}$$

$$= \begin{cases} (1-p) \cdot \frac{1}{n} + p \cdot \frac{w_{ij}}{c_j} & \text{si} \quad c_j \neq 0 \\ \frac{1}{n} & \text{si no} \end{cases}$$

pero:

$$a_{ij} := Pr(j \longrightarrow i) = \begin{cases} (1-p) \cdot \frac{1}{n} + p \cdot \frac{I_{ij}}{c_j} & \text{si} \quad c_j \neq 0 \\ \frac{1}{n} & \text{si no} \end{cases}$$

Como $I_{ij} = 1$ si y sólo si existe un hipervínculo de j a i, con $j \neq i$ —y nulo en caso contrario—, entonces $I_{ij} = w_{ij}$ y concluímos que $a_{ij} = (p\mathbf{WD} + ez^t)_{ij}$, $\forall i, j : 1 \ldots n$, lo que implica que:

$$\mathbf{A} = p\mathbf{WD} + ez^t$$

■

5.2. **$\mathbf{I} - p\mathbf{WD}$ permite la eliminación gaussiana.**

Referencias

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30 (1-7):107–117*, 1998.

[2] Kurt Bryan and Tanya Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *T SIAM review, 48 (3):569–581*, 2006.

[3] Sepandar D Kamvar; Taher H Haveliwala; Christopher D Manning and Gene H Golub. Extrapolation methods for accelerating pagerank computations. *In Proceedings of the 12th international conference on World Wide Web, pages 261-270. ACM*, 2003.