Centro de Investigación y Docencia Económicas

Algoritmo de Búsqueda de Google *PageRank*

Jair Ivan Garcia Gonzalez

José Raúl Félix Martínez

Sara Briseño Cruz

Sergio Eduardo Tovar Benítez

Raciel Vasquez Aguilar

Matemáticas III

24 de Noviembre 2021

Índice

1 Introducción

1.1 Antecedentes

La búsqueda y recuperación de información es el proceso por el cual se busca una información en particular dentro de una colección de documentos.[1]

Los primeros sistemas de búsqueda computarizados utilizaban un sistema que automáticamente recuperaba información de documentos relacionados con la consulta del usuario. Sin embargo, se limitaba mayormente al uso de bibliotecarios.

En 1989 la creación de la World Wide Web revolucionó el almacenamiento, acceso y búsqueda de colecciones de documentos.[2] A pesar de los avances en estos ámbitos, los usuarios que comenzaban a utilizar la búsqueda web eran abrumados por la cantidad de información entre la que tenían que buscar y la que era de su interés. Por ello, muchos usuarios se guiaban según las recomendaciones de conocidos y consejos de expertos.

Con la creación del análisis de enlaces en 1998,[3] los motores de búsqueda empezaron a implementar esta técnica que utiliza la información adicional que brindan los hipervínculos para mejorar la calidad de los resultados de búsqueda.

1.2 PageRank

Diseñado en 1998 por Sergei Brin y Lawrence Page, estudiantes de doctorado en informática de la universidad de Stanford, el sistema de recuperación de información de Google PageRank, es un algoritmo basado en las cadenas de Markov y sirve para recuperar páginas web en línea tomando en cuenta el interés humano y la atención prestada.

En la actualidad, Google se ha convertido en el motor de búsqueda dominante a nivel mundial gracias a su sistema independiente de consulta, a la inmunidad que ha logrado

[1] Amy N. Langville y Carl D. Meyer, *Google's PageRank and beyond: The science of Search Engine Rankings,* (New Jersey: Princeton University Press, 2012), PDF

[2] James Gillies and Robert Cailliau, *How the Web Was Born: The Story of the World Wide Web*, (Oxford: Oxford University Press, 2000), PDF.

[3] Amy N. Langville y Carl D. Meyer, *Google's PageRank and beyond: The science of Search Engine Rankings.* (New Jersey: Princeton University Press, 2012.), PDF

contra el *spamming* y el éxito del negocio de Google.[4] Debido a la carga política que tiene el ofrecer información a una gran cantidad de personas, Google tiene un problema técnico y jurídico en el que debe satisfacer la demanda de información veraz y pertinente a los usuarios a partir de datos imposibles de revisar manualmente por su exuberante cantidad. Para 2002, la cantidad de páginas web accesibles desde la plataforma era de 2.7 billones.[5]

Además, debido a las potencialidades legales y económicas de internet, todos los usuarios tienen incentivos para aumentar artificialmente la cantidad de visitas que obtiene determinada página web.

En este contexto, el presente trabajo se plantea el modelo de PageRank utilizado anteriormente por Google y desactualizado. Por razones de seguridad, Google no publica los algoritmos que utiliza para priorizar un sitio web sobre otro en una búsqueda cualquiera. Sin embargo, es posible conjeturar que los modelos actuales no distan mucho del PageRank aquí presentado y que los cambios hechos por seguridad son adiciones al concepto general de 1998.

2 Desarrollo teórico

2.1 Teoría de Grafos

Un grafo es un par ordenado de dos conjuntos: nodos y vínculos, denotados por $G = (N, V)$. Un vínculo llamado xy conecta los nodos x y y. El orden de un grafo es igual a la cantidad de nodos existentes.[6]

En un grafo, un camino C es una secuencia donde se alternan nodos y vínculos de

[4] Amy N. Langville y Carl D. Meyer, *Google's PageRank and beyond: The science of Search Engine Rankings.* (New Jersey: Princeton University Press, 2012.), PDF

[5] "The World's Largest Matrix Computation", Cleve Moler, MathWorks, última modificación: 2002, consultado el 20 de noviembre de 2021, https://la.mathworks.com/company/newsletters/articles/the-world-s-largest-matrix-computation.html

[6] Juan Manuel Barriola y Milena Dotta, "¿Cómo funciona Google? El algoritmo PageRank, Diagramas de grafos y cadenas de Markov," Revista de Investigación en Modelos Matemáticos Aplicados a la Gestión y la Economía Año 3 – N° 3 (2016): 11, Latindex.

la siguiente forma:

$$x_0, v_1, x_1, v_2, ..., v_n, x_n \quad \text{con} \quad v_i = x_{i-1}x_i 0 < i \leq n$$

La longitud del camino se define a partir de n. No hay restricción sobre que nodos conforman un camino, por lo que pueden repetirse.

Un caso particular de un camino C es un sendero S, en este se excluyen los casos de recurrencia a un mismo nodo, es decir, una vez que se ha pasado por cierto nodo no se vuelve a pasar por él.

Carácter restringido entre los nodos. Un vínculo dirigido se define como aquel que brinda una relación de dirección entre los nodos. Se suele denotar como $\overrightarrow{xy}$ a un vínculo que parte de x hacia y. De modo que, el nodo desde el cual parte el vínculo se denomina *nodo de salida* y el vértice al que llega se llama *nodo de llegada.* Un grafo orientado es aquel en el que todos los vínculos que lo componen tienen una orientación definida. Es decir, si se tienen los nodos xy, debe verificarse al menos que $\overrightarrow{xy}$ o $\overrightarrow{yx}$.

Transiciones entre nodos. Un grafo ponderado es un tipo de grafo orientado donde a cada vínculo se le asigna un valor no negativo llamado ponderación. Entonces, sea v_i un vínculo, $P(v_i)$ será su ponderación.

Vale la pena hacer las siguientes distinciones en relación a las conexiones entre nodos: un grafo orientado está conectado si para todo par de nodos $\{x, z\}$ existe un *camino dirigido* desde x hasta z, **o** desde z hasta x, por lo que existen caminos entre todos los nodos. Mientras que, un grafo orientado está *fuertemente* conectado si para todo par de nodos $\{x, z\}$ existe un *camino dirigido* desde x hasta z, **y** desde z hasta x.

2.2 Cadenas de Markov

Matriz estocástica. Es una matriz no negativa P_{nxn} en que la suma de cada fila es igual a 1. Por ejemplo:

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

Proceso estocástico. Conjunto de variables aleatorias $\{X\}_{t=0}^{\infty}$ con un rango común que es llamado espacio de estados. El parámetro t es pensado como el tiempo y X_t representa el estado del proceso en el instante t. Cadenas de Markov. Es un proceso estocástico que e satisface la propiedad de Markov:

$$P(X_{t+1} = S_j | X_t = S_{i_t} = S_{i_{t-1}}, ..., X_0 = S_{i0}) = P(X_{t+1} = S_j | X_t = S_{i_t})$$

Probabilidad de transición y matriz. Es la probabilidad de pasar del estado S_i en el instante t-1 al estado S_j en el instante t. Y su matriz P_{nxn}(t)=$[p_{ij}(t)]$ es una matriz no negativa y la suma de cada fila es 1.En otras palabras, P(t) es una matriz estocástica para cada t.

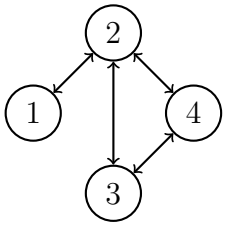
Vector de distribución de probabilidad. Se define como un vector fila no negativo $P^T = (p_1, p_2, ..., p_n)$, tal que $\sum_k Pk = 1$

2.3 Algoritmo PageRank

La red de información en internet puede modelarse como un multígrafo o un grafo donde los vértices son páginas y las aristas son hipervínculos que conectan a una página con otra. Una página está conectada a otra si hay un hipervínculo en esa página que la conecte con la otra. Una misma página puede tener más de un hipervínculo con otra,

este hecho hace que el modelado de la red de información sea un multígrafo.[7]

En este sentido, conviene considerar una interpretación alternativa, matricial. Sea M de dimensiones nxn, cuyas entradas son ceros y unos. La entrada m_{ij} de la matriz será un uno si es que hay un enlace de la página P_j a la página P_i y un cero en caso contrario. Para el siguiente grafo, podemos formar dos matrices de adyacencia:



$$MatrizA \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad MatrizB \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

En términos simples, Google utiliza el grafo de las páginas de internet para modelar una votación ponderada de los sitios más relevantes y así, puntuar la relevancia de un determinado sitio web. La matriz utilizada para PageRank tiene traza 0, como en la matriz A, pues se determina que la relevancia de una página web no aumenta por estar conectada a sí misma. Sin embargo, para otras aplicaciones, la matriz B puede ser de utilidad y mantiene la estructura del grafo correspondiente invariante.

Es importante hacer la distinción entre dos grafos con aristas unidireccionales y bidireccionales. Una página puede referenciar y ser referenciada por otras páginas, pero ambos casos son independientes.

Tomando la matriz de adyacencia, la suma de los elementos de determinada columna es igual al número de páginas que hacen referencia a la página correspondiente. Esta suma establece la cantidad de páginas que apuntan a determinado sitio web. Desde el punto de vista del cálculo es muy fácil obtener la matriz P a partir de la matriz A. Formalizando este hecho, definamos el número de enlaces salientes de una página j como: $c_j = \sum_{i=1}^{n} g_{ji}$, $1 \leq j \leq n$.

Si se ponderan los valores de las columnas de la matriz de adyacencia, por el número de dependencias de cada sitio web, obtenemos una matriz de fracciones cuya suma es

[7]Espinoza Armenta, *Matemáticas Discretas.* (México: Alfaomega, 2010), PDF.

1. Formalizando este hecho, la matriz $P = (p_{ij}) \in R^{nxn}$, tal que:

$$p_{ij} = \begin{cases} g_{ij}/c_j & \text{si } c_j \neq 0 \\ 0 & \text{En otro caso} \end{cases}$$

Es obvio que si $c_j \neq 0$, para todo j, entonces P es una matriz estocástica por columnas, es decir, la suma de cada columna vale 1 y cada elemento toma un valor entre 0 y 1. Tomemos el siguiente ejemplo:

$$Matriz \quad A \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad A \quad ponderada \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

La matriz A, además de cumplir con lo anterior, tiene otras propiedades matemáticas interesantes: su radio espectral es uno, la matriz A es irreducible y primitiva. Estas propiedades son importantes ya que el vector PageRank es el vector límite de distribución de probabilidad de una cadena de Markov. En el ejemplo anterior la matriz A es primitiva, pero hay casos en los que la estructura de enlaces entre las páginas no conduce a una matriz estocástica y primitiva.

La velocidad de procesamiento y acceso y el almacenamiento de los datos de la red no son tareas triviales. ¿Por qué no utilizar las sumas columna y fila para obtener un PageRank simple y aritmético? Muy simple, porque es posible crear automáticamente cientos y miles de sitios que hagan referencia a determinada página web y agregarlos como referencia en esa misma página. Sin embargo, la importancia de los sitios automáticos, sin visitas y no referenciadas por otras direcciones es nula en PageRank y la relevancia de cualquier página web depende del estado general de todas las demás direcciones de internet.

En esta linea, el problema ahora consiste en decidir que la importancia x_j de cada pagina P_j es proporcional a la suma de las importancias de las paginas que enlazan con

P_j. El cambio que hemos hecho, respecto a la premisa del principio, consiste en que se cambio "número de páginas que enlazan con P_j" por "la suma de las importancias de las paginas que enlazan con P_j".

Propongamos un ejemplo. Supongamos que la página P_1 es citada desde las páginas P_2, P_{25} y P_{256}, que P_2 sólo se cite desde P_1 y P_{256}, etc., mientras que, digamos, hay enlaces a la última página, P_n, desde $P_1, P_2, P_3, P_{25} y P_{n-1}$. En nuestra asignación anterior, x_1 debería ser proporcional a 3, x_2 lo sería a 2, etc., mientras que x_n habría de ser proporcional a 5. Pero ahora nuestra asignación $x_1, ..., x_n$ debe cumplir que:

$$x_1 = K(x_2 + x_{25} + x_{256}),$$

$$x_2 = K(x_1 + x_{256})$$

$$x_n = K(x_1 + (x_2 + x_3 + x_{25} + x_{256})$$

Donde K es una cierta constante de proporcionalidad. Ahora, hay un sistema de ecuaciones, cuyas soluciones son las posibles asignaciones $x_1, ..., x_n$. Hagamos un pequeño cambio, llamemos x al vector de importancias. La matriz de dimensiones nxn del sistema es, justamente, la M asociada al grafo y la constante de proporcionalidad la cambiemos por λ. Así, podremos escribir que la asignación de importancias que perseguimos es una solución de:

$$Mx = \lambda x$$

2.4 Modelo de Brin y Page

El modelo inicial para el cálculo del vector PageRank se basaba en calcular el vector estacionario de la matriz P de orden n, siempre que esta matriz fuera estocástica y primitiva. En este modelo no se contemplan los nodos sin salida con lo cual cj es no nulo para todo j y, en consecuencia, P es estocástica. Sin embargo, se dieron cuenta que la estructura de la web daba lugar a que P no fuera primitiva e introdujeron un nuevo modelo basado en una matriz estocástica F definida de la siguiente manera:

$$F = \alpha P + (1 - \alpha) v e^T$$

donde $0 < \alpha < 1, e^T \in R^{1xn}$ es el vector de unos y v $\in R^{nx1}$ es el llamado vector de personalización. V es un vector de distribución de probabilidad que se suele tomar como v $= \frac{1}{n} e$. El parámetro α se denomina de amortiguamiento (damping) y se suele tomar α= 0.85, ya que fue el que usaron originalmente Brin y Page. El termino $(1 - \alpha) v e^T$, da lugar a que todos los elementos de F sean no nulos, con lo cual es irreducible.

Ahora, cuando hay nodos que no tienen enlaces salientes en las columnas respectivas se tiene que $c_j = 0$ y estas columnas están llenas de ceros. En consecuencia P no será estocástica ni tampoco lo será F. Entonces, definimos una nueva matriz:

$$F' = \alpha(P + v d^T) + (1 - \alpha) v e^T$$

Donde lo único que cambia es el vector $d \in R^{nx1}$, definido como $d_i = 1$, si $c_i = 0$, y $d_i = 0$ en otro caso. De esta forma la matriz es estocástica aunque haya nodos sin salida. La matriz $v d^T$ actúa sobre las columnas pertenecientes a nodos sin salida, asignándoles una probabilidad no nula. La matriz F' tiene la desventaja de ser muy densa y enorme, pero como se explica en la sección siguiente no hace falta calcularla explícitamente.

2.5 El método de la potencia

El elegido por Google para calcular PageRank. Es una técnica iterativa que calcula aproximaciones sucesivas a los vectores propios y valores propios de una matriz $A \in R^{mxm}$ diagonalizable con autovalores.

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ...|\lambda_n| \geq 0$$

En comparación con otros métodos iterativos este suele ser el más lento y solo puede determinar uno de los vectores propios de la matriz.

El método de la potencia puede resumirse en el siguiente **teorema:** Sea x_o arbitrario,

si definimos

$$y_n = Ax_n, v_n = m(y_n), x_{n+1} = \frac{y_n}{v_n}, \quad \text{para} \quad n = 0, 1, 2, ..., n(1)$$

Entonces $x_n \to x$ y $v_n \to \lambda_1$, donde $Ax = \lambda_1 x$

A pesar de sus inconvenientes ya mencionados, el método de la potencia resulta ideal para calcular el vector PageRank, de acuerdo con Ángela Piñeiro Lourés,[8] las razones matemáticas son:

1. Cada iteración requiere sólo de un producto matriz-vector, y esto puede ser aprovechado para reducir el esfuerzo computacional cuando A es grande y dispersa.

2. Los cálculos pueden ser hechos en paralelo mediante el cómputo simultáneo de productos internos de filas de A con x_n.

3. Para una matriz diagonalizable, el ritmo al que converge (1) depende de qué tan rápido $(\frac{\lambda_2}{\lambda_1})^n \to 0$. Se sabe que Google es capaz de controlar el ritmo de convergencia ya que puede regular $|\lambda_2|$.

4. Las iteraciones son $x_{n+1} = Ax_n$ ya que para el problema PageRank $\lambda_1 = 1$.

De igual forma, Amy Langville y Carl Meyer[9] explican que existen diversas razones computacionales por las cuales era buena idea utilizar el método de la potencia:

1. Al ser un método tan simple su implementación y programación es sencilla.

2. Es lo que se conoce en matemáticas computacionales, como un método sin matrices. Esto significa que la matriz no se manipula, y ya que modificar y almacenar elementos de la matriz de Google no es factible por su tamaño, el método de la potencia es preferido.

[8] Angela Piñero Lourés, *Las matemáticas del algoritmo PageRank.* (Galicia: Universidad de Santiago de Compostela, 2020), PDF

[9] Amy N. Langville y Carl D. Meyer, *Google's PageRank and beyond: The science of Search Engine Rankings.* (New Jersey: Princeton University Press, 2012.), PDF

3. Es óptimo para el almacenamiento ya que se debe almacenar solamente un vector con n números reales y en el caso de Google $n = 8.1$ mil millones.

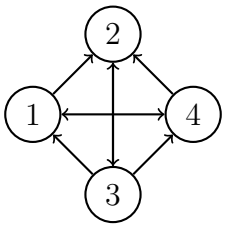
4. El número de iteraciones. Brin y Page reportaron que son necesarias entre 50 y 100 iteraciones antes de que converjan, dando una aproximación satisfactoria al vector exacto de PageRank.

3 Ejemplos desarrollados

3.1 Ejemplo con 4 páginas

Consideremos una web en la que solo existen cuatro páginas. Cada una representada por un nodo. Los hipervínculos son representados mediante las flechas que a su vez corresponden a vínculos dirigidos.

En la siguiente figura se muestra el grafo dirigido G.



La matriz de adyacencia y su correspondiente ponderada son:

$$Matriz \quad A \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \qquad A \ ponderada \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

$$A \ \ ponderada \ \ transpuesta \begin{bmatrix} 0 & 0 & 1/3 & 1/2 \\ 1/2 & 0 & 1/3 & 1/2 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 \end{bmatrix}$$

Calculamos los valores y vectores propios

Factor:

$$(\lambda - 1)(\lambda + \frac{1}{2})(\lambda^2 + \frac{1}{2}\lambda + \frac{1}{6})$$

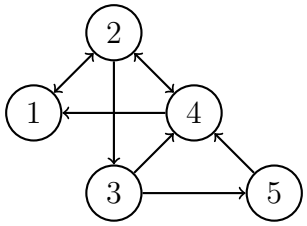
Valores Propios:

$$\lambda_1 = 1, \lambda_2 = -\frac{1}{2}, \lambda_3 = \frac{-i\sqrt{15} - 3}{12}, \lambda_4 = \frac{i\sqrt{15} - 3}{12}$$

Vector propio de $\lambda = 1$:

$$v_1 = (1, \frac{3}{2}, \frac{3}{2}, 1)$$

3.2 Ejemplo con 5 páginas



La matriz de adyacencia transpuesta y su correspondiente ponderada son:

$$Matriz \quad A^{\,t} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad A^{\,t} \quad ponderada \begin{bmatrix} 0 & 1/3 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 & 0 \end{bmatrix}$$

Considerando que $Ax = \lambda_1 Ix$, podemos obtener un sistema de ecuaciones lineales homogéneo donde $(A - \lambda_1 I)x = 0$. Por tanto, la forma matricial de $(A - \lambda_1 I)$ queda de la siguiente forma:

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

De tal forma, el determinante de la matriz $(A - \lambda_1 I)$ corresponde a:

$$P(\lambda) = -\lambda^5 + \frac{1}{2}\lambda^3 + \frac{1}{4}\lambda^2 + \frac{1}{6}\lambda + \frac{1}{12} = 0$$

Donde $\lambda_1 = 1$

$$(A - \lambda_1 I)x = \begin{bmatrix} -1 & 1/3 & 0 & 1/2 & 0 \\ 1 & -1 & 0 & 1/2 & 0 \\ 0 & 1/3 & -1 & 0 & 0 \\ 0 & 1/3 & 1/2 & -1 & 1 \\ 0 & 0 & 1/2 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

Al multiplicar las matrices, obtenemos el siguiente sistema de ecuaciones:

$$\begin{aligned} -x_1 + \frac{1}{3}x_2 + \frac{1}{2}x_4 &= 0 \\ x_1 - x_2 + \frac{1}{2}x_4 &= 0 \\ \frac{1}{2}x_2 - x_3 &= 0 \\ \frac{1}{3}x_2 + \frac{1}{2}x_3 - x_4 + x_5 &= 0 \\ \frac{1}{2}x_3 - x_5 &= 0 \end{aligned}$$

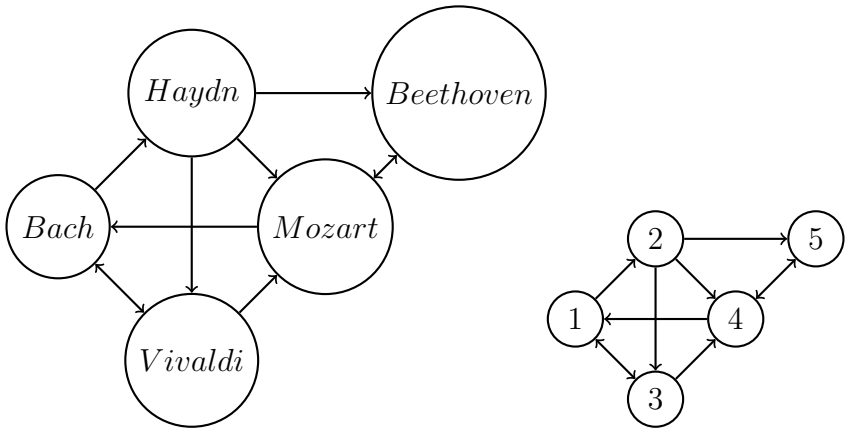
De esta manera, obtenemos el vector que corresponde al ranking de las páginas:

$$[4, 6, 2, 4, 1]$$

4 Problemas aplicados resueltos

En wikipedia hay 5 artículos relacionados que pueden representarse con un subgrafo dirigido. El artículo sobre Bach apunta a Haydn y Vivaldi, Haydn referencia a Vivaldi, a Beethoven y a Mozart, Vivaldi referencia a Bach y a Mozart, Bethoven referencia a Mozart y Mozart referencia a Bach y a Beethoven.

En la siguiente figura se muestra el grafo dirigido del sistema.



La matriz de adyacencia y su correspondiente ponderada son:

$$Matriz \quad A \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad A \quad ponderada \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Utilizamos la matriz ponderada transpuesta

$$A\,t \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 1/2 & 0 & 1 \\ 0 & 1/3 & 0 & 1/2 & 0 \end{bmatrix} \rightarrow \quad \det \begin{bmatrix} -\lambda & 0 & 1/2 & 1/2 & 0 \\ 1/2 & -\lambda & 0 & 0 & 0 \\ 1/2 & 1/3 & -\lambda & 0 & 0 \\ 0 & 1/3 & 1/2 & -\lambda & 1 \\ 0 & 1/3 & 0 & 1/2 & -\lambda \end{bmatrix}$$

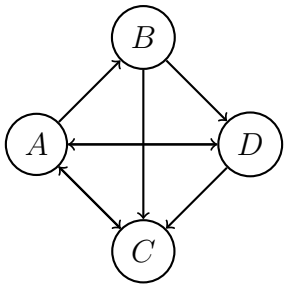
Los valores propios de la matriz $A\,t$ son $\lambda_1 = 1;\ \lambda_2 = 0.29;\ \lambda_3 = -0.67;$
$\lambda_4 = -0.31 + 0.34i;\ \lambda_5 = -0.31 - 0.34i;$

El vector propio asociado a λ_1 es $\begin{bmatrix} 0.48 & 0.24 & 0.32 & 0.65 & 0.40 \end{bmatrix}$. Por lo que al presentar las páginas de wikipedia de un qwery sobre compositores occidentales del siglo XVIII, el orden de aparición sería Mozart, Bach, Beethoven, Vivaldi y Haydn. En este ejercicio no se incluye la constante normalizadora, pero el resultado final sería el mismo aunque se incluyese.

La magnitud del orden n de la matriz de adyacencia de Google, ronda los ocho billones. Crear un grafo arbitrario de ocho billones de vértices, obtener su matriz de adyacencia y encontrar el vector PageRank de la misma, se deja como ejercicio reto para el lector.

4.1 4 páginas de igual importancia

Volvamos a Considerar una web en la que solamente existen cuatro páginas, pero a diferencia del ejemplo 3.1, en este caso agregamos la ponderación que le corresponde a cada caso, suponiendo que la importancia esta distribuida de manera equitativa entre las cuatro páginas.



Se puede notar que el grafo se encuentra fuertemente conectado. Debido a esto la

matriz adyacente es estocástica.

$$Matriz \quad Adyacente \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Sea v el vector de ranking inicial y ya que inicialmente la importancia está distribuida de manera uniforme, tenemos que $v = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Para actualizar la importancia se multiplica la matriz adyacente por el vector v. El resultado es un nuevo vector de importancia nombrado $v_1 = vA$. Al repetir el proceso se llega a un nuevo vector de importancia denominado $v_2 = (vA)A = vA^2$. Si reiteramos infinitamente este proceso, se converge al vector de importancia de equilibrio.

En este ejemplo sería:

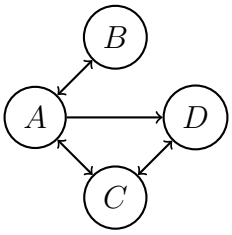
iteración	vectores
v	$(0.25, 0.25, 0.25, 0.25)$
vA	$(0.37, 0.08, 0.33, 0.20)$
vA^2	$(0.43, 0.12, 0.27, 0.16)$
vA^3	$(0.35, 0.14, 0.29, 0.20)$
vA^4	$(0.39, 0.11, 0.29, 0.19)$
vA^5	$(0.39, 0.13, 0.28, 0.19)$
vA^6	$(0.38, 0.13, 0.29, 0.19)$
vA^7	$(0.38, 0.12, 0.29, 0.19)$
vA^8	$(0.38, 0.12, 0.28, 0.19)$
vA^9	$(0.38, 0.12, 0.29, 0.19)$
vA^{10}	$(0.39, 0.13, 0.29, 0.19)$

La interpretación de los resultados indican que de acuerdo a este algoritmo la página A será la primera en el ranking por su relevancia a la búsqueda, seguida por la C. Quedando las páginas D y B en tercer y cuarto puesto respectivamente.

5 Ejercicios con respuestas

5.1 Ejercicio con 4 páginas

A partir de la siguiente figura, determine el vector correspondiente a la quinceava iteración, y, por consiguiente, el ranking de las páginas A, B, C y D según su relevancia en la búsqueda.



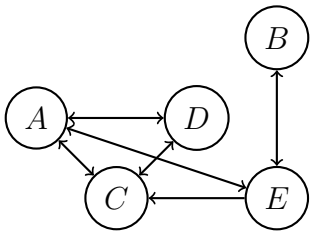
Respuesta:

$$vA^{15} = (0.2722, 0.09117, 0.3637, 0.2728)$$

Por tanto, la página D ocupa el primer puesto en el ranking, seguido de la página A y C que ocupan el segundo y tercer puesto correspondientemente, dejando al final a la página B.

5.2 Ejercicio con 5 páginas

De acuerdo a la siguiente figura, plantee el sistema de ecuaciones lineales homogéneo y obtenga el vector asociado al ranking de las páginas A, B, C, D, E y F según su relevancia en la búsqueda.



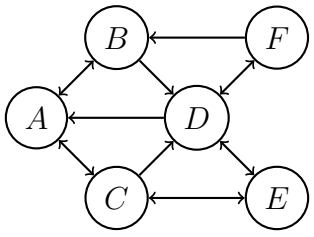
Respuesta:

$$\begin{aligned} -x_1 + \frac{1}{2}x_3 + \frac{1}{2}x_4 + \frac{1}{3}x_5 &= 0 \\ -x_2 + \frac{1}{3}x_5 &= 0 \\ \frac{1}{3}x_1 - x_3 + \frac{1}{2}x_4 + \frac{1}{3}x_5 &= 0 \\ \frac{1}{3}x_1 + \frac{1}{2}x_3 - x_4 &= 0 \\ \frac{1}{3}x_1 + x_2 - x_5 &= 0 \end{aligned}$$

El vector correspondiente al ranking de las páginas es $(2, \frac{1}{3}, \frac{16}{9}, \frac{14}{9}, 1)$.

5.3 Ejercicio con 6 páginas

A partir de la siguiente figura, encuentre el ranking de las páginas A, B, C, D, E y F, según su relevancia en la búsqueda, con el método de su preferencia.



Respuesta: El ranking de las páginas es el siguiente: D, A, E, C, B, F.

6 Referencias

Langville, Amy N., y Carl D. Meyer. *Google's PageRank and beyond: The science of Search Engine Rankings.* New Jersey: Princeton University Press, 2012. PDF.

Gillies, James, y Robert Caillau. *How the Web Was Born: The Story of the World Wide Web.* Oxford: Oxford University Press, 2000. PDF.

Barriola, Juan Manuel y Milena Dotta. "¿Cómo funciona Google? El algoritmo PageRank, Diagramas de grafos y cadenas de Markov." Revista de Investigación en Modelos Matemáticos Aplicados a la Gestión y la Economía Año 3 – N° 3 (2016): 9-30. Latindex.

Piñero Lourés, Ángela. *Las matemáticas del algoritmo PageRank.* Galicia: Universidad de Santiago de Compostela, 2020. PDF.

Fernández, Pablo. "El secreto de Google y el álgebra lineal." Boletín de la Sociedad Española de Matemática Aplicada n°30(2004), 115-141. Boletín electrónico SEMA.

Armenta, Espinoza. *Matemáticas Discretas.* México: Alfaomega, 2010. PDF.