¿CÓMO FUNCIONA GOOGLE? EL ALGORITMO PAGERANK, DIAGRAMAS DE GRAFOS Y CADENAS DE MARKOV

JUAN MANUEL BARRIOLA y MILENA DOTTA

Facultad de Ciencias Económicas. UBA
Av. Córdoba 2122 – 2do piso
C1120AAQ – Ciudad Autónoma de Buenos Aires. Argentina

juanbarriola@hotmail.com mile.dotta@hotmail.com

Recibido 10 de mayo de 2015, aceptado 3 de junio de 2015

Resumen

Los motores de búsqueda en Internet han evolucionado considerablemente en el último tiempo. En el presente trabajo se pretende explicar el funcionamiento del algoritmo PageRank de Google en su versión más elemental.

Se utiliza la teoría de grafos para brindar una representación de la red de páginas de Internet con diagramas de grafos y sus matrices asociadas. Por otra parte, al concebir la búsqueda web como un fenómeno aleatorio, la misma se puede abordar mediante los conceptos de Cadenas de Markov.

Al confirmar que la matriz de adyacencia de un grafo fuertemente conectado comparte las características de una matriz estocástica se demuestra que el algoritmo asigna la importancia de las páginas de interés para el navegante iterando infinitas veces la misma búsqueda, es decir, calcula el vector punto fijo de la matriz.

Se muestran los límites de esta versión del algoritmo al nombrar dos características que puede presentar la red de páginas de Internet. Se concluye con un ejemplo numérico de lo expuesto.

Palabras clave: Motores de búsqueda - Diagramas de grafos- Cadenas de Markov - Vector punto fijo.

HOW DOES GOOGLE WORK? PAGERANK ALGORITHM, GRAPH DIAGRAMS AND MARKOV CHAINS

JUAN MANUEL BARRIOLA Y MILENA DOTTA

Facultad de Ciencias Económicas. UBA
Av. Córdoba 2122 – 2do piso
C1120AAQ – Ciudad Autónoma de Buenos Aires. Argentina
juanbarriola@hotmail.com mile.dotta@hotmail.com

Abstract

In the past few years web search engines have evolved increasingly faster. This article seeks to explain the most basic way Google's PageRank algorithm works.

Graph theory is utilized to represent the web of Internet pages with graph diagrams and its associated matrices. Secondly, we show that if web search is conceived as a random walk, it can be analyzed using concepts related to Markov Chains.

After proving that the adjacency matrix of a strongly connected graph has the characteristics of an stochastic matrix it is shown that the importance of the pages is assigned by the algorithm by repeating infinites times the same search, namely, it calculates the matrix's fixed probability vector.

The limits of this basic version of the algorithm are shown by exposing two usual features the structure of the web may present. We conclude with a numerical example of the concepts previously exposed

Keywords: Web search engines – Graph diagrams - Markov Chain - Fixed probability vector

INTRODUCCIÓN

En el presente trabajo nos proponemos realizar una explicación básica de los algoritmos de búsqueda de Internet más comunes y usados en la actualidad. Se pretende demostrar la ventaja que este tipo de algoritmos suponen en comparación a otros métodos más antiguos, aquellos en los cuales la búsqueda se realiza a partir de los textos: Text Based Ranking Systems.

Se introducirán conceptos básicos de Teoría de grafos que son utilizados para representar a la red de páginas de Internet. A su vez, serán explicitados ciertos conceptos referidos a las Cadenas de Markov ya que son relevantes al modelar la búsqueda en páginas web como un fenómeno aleatorio.

Con ambos desarrollos, será posible explicar de manera muy elemental cómo es que funciona el algoritmo de los motores de búsqueda más populares.

Por último se presentará un ejemplo numérico para facilitar la comprensión de los temas expuestos en las secciones precedentes.

1. TEORÍA DE GRAFOS

En esta sección se exponen los conceptos elementales de teoría de grafos considerados como necesarios para proseguir en la explicación en las secciones subsiguientes.

Un grafo es un par ordenado de dos conjuntos: nodos y vínculos, que se denota $G = (N, V)$. Un vínculo $\{xy\}$ se dice que conecta a los nodos *x* e *y,* y es denominado *xy*. El orden de un grafo es igual a la cantidad de nodos existentes (Biggs, 1993) (Beineke, Wilson, Cameron, 2005) (Bollobás, 1995).

Un camino *C* en un grafo es una secuencia que alterna nodos y vínculos de la siguiente manera:

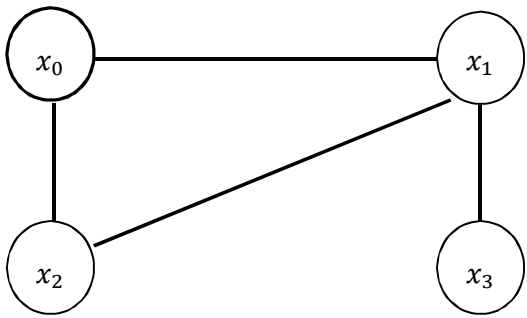
$$x_0, v_1, x_1, v_2, \dots, v_l, x_l \qquad con \qquad v_i = x_{i-1}x_i \quad 0 < i \leq l$$

La longitud del camino se define a partir de *l*. No se impone ninguna restricción sobre cuáles son los nodos que conforman un camino, permitiendo que se repitan nodos en un mismo camino.

Un sendero *S* es un caso particular, más restrictivo, de camino ya que se excluyen los casos de recurrencia a un mismo nodo, es decir, una vez que se ha pasado por determinado nodo no se vuelve a pasar por él. (Beineke, Wilson, Cameron, 2005)

La potencia y utilidad de los grafos se debe a que por las definiciones antes expuestas es natural representarlos gráficamente. Contando así con una descripción clara y sencilla para interpretar lo expresado mediante notación formal.

Grafo 1. Grafo simple



Hasta ahora, mediante los conceptos presentados, se ha llegado a definir un grafo en su manera más general. Pero a los intereses del presente trabajo es necesario introducir definiciones que permitan describir una

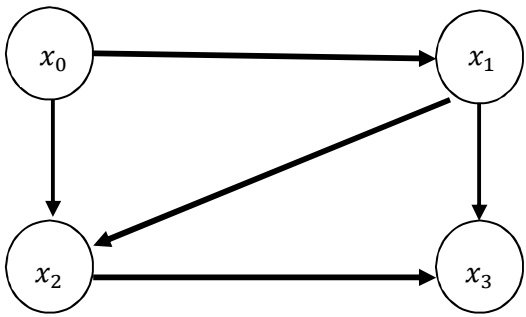
clase de grafos en los que existan relaciones de carácter más restringido entre los nodos.

Para ello se parte de definir a un vínculo dirigido como aquel que brinda una relación de dirección entre los nodos. Considerando a un vínculo que surge desde *x* hacia *y,* se lo suele denotar como $\overrightarrow{xy}$ para diferenciarlo de los vínculos corrientes. De esta manera, el nodo desde el cual parte el vínculo se denomina *nodo de salida* y hacia dónde llega se llama *nodo de llegada.*

Así, un grafo orientado se define como aquel en el que todos los vínculos que lo componen tienen una orientación definida. Teniendo dos nodos *x y* cualesquiera, debe verificarse que debe existir como mínimo un vínculo dirigido entre ellos, es decir, se cumpla al menos que $\overrightarrow{xy}$ o $\overrightarrow{yx}$. (Bollobas, 1995)

Los conceptos de camino y sendero son aplicables para un grafo orientado, al ser este una forma particular de grafo. En este caso, los caminos y senderos son orientados. En cada uno de ellos, el conjunto de vínculos que lo componen debe respetar la orientación de los mismos.

Grafo 2. Grafo orientado



Los vínculos ahora permiten definir el sentido en el que se produce el movimiento de un nodo a otro. Es necesario avanzar en una caracterización

más precisa, que brinde una mayor información sobre las transiciones entre nodos.

Un grafo ponderado es un grafo orientado en el que a cada vínculo se le asigna un valor no negativo denominado ponderación. Sea v_i un vínculo, $p(v_i)$ es su ponderación.

Aunque la definición no plantea ninguna restricción sobre el valor que puede tomar la ponderación de un grafo, usualmente se trabaja con medidas normalizadas. En el caso en el cual sólo exista un vínculo partiendo de determinado nodo, su ponderación será igual a uno. En el marco de este trabajo se trabajará con dicha normalización, por lo tanto debe cumplirse:

$$0 < p(v_i) \leq 1 \qquad \forall v_i \in V(G)$$

Esto a su vez indica que siendo *x* un nodo de salida, v_i los vínculos que surgen desde *x* y $p(v_i)$ las ponderaciones asociadas, debe cumplirse que:

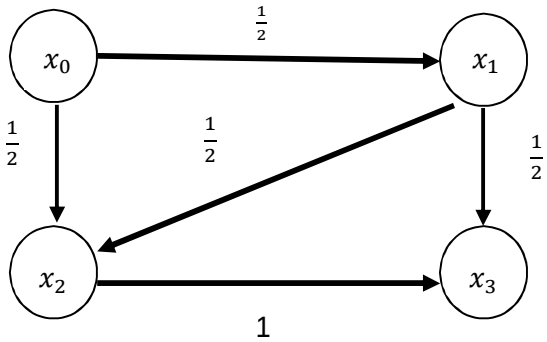
$$\sum_{i=1}^{I} p(v_i) = 1$$

La sumatoria de las ponderaciones de todos los vínculos que surgen desde el mismo nodo, es decir, que comparten al nodo de salida, debe ser igual a uno.

Existen dos formas de considerar las ponderaciones asociadas a los vínculos:

- Forma equitativa: Las ponderaciones de los vínculos dirigidos desde el nodo *x* a cualquier nodo (incluido sí mismo) son idénticas. Por lo tanto, si existen *i* vínculos saliendo desde *x* entonces $p(v_i) = \frac{1}{i}$
- Forma arbitraria: Las ponderaciones de los vínculos son diferentes. En este sentido, se considera que cuanto mayor sea la ponderación de un vínculo respecto a los restantes, mayor será su importancia.

Grafo 3. Grafo ponderado



Esta misma información que se encuentra en la representación del grafo puede utilizarse para construir una matriz que será de suma utilidad.

Adaptando la definición que brinda Bollobás para grafos no ponderados se plantea:

Sea G un grafo ponderado con *n* nodos, la matriz de adyacencia (A) asociada al grafo es una matriz de orden *n x n* cuyos elementos se construyen de la siguiente manera:

$$a_{xy} = \begin{cases} p(\overrightarrow{xy}) & si \quad x\ e\ y\ son\ nodos\ adyacentes \\ 0 & si \quad x\ e\ y\ no\ son\ nodos\ adyacentes \end{cases}$$

$$A = \begin{pmatrix} a_{x_1x_1} & \cdots & a_{x_1x_n} \\ \vdots & \ddots & \vdots \\ a_{x_nx_1} & \cdots & a_{x_nx_n} \end{pmatrix}$$

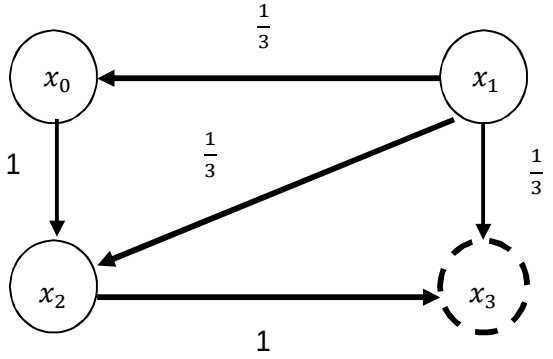
Con las consideraciones realizadas sobre las ponderaciones, necesariamente se obtiene que:

- $0 \leq a_{xy} \leq 1 \quad \forall\, x, y \in G(N.V)$

Para finalizar, se introducen dos distinciones más sobre los grafos en relación a las conexiones entre nodos.

Se dice que 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.* (Bollobás, 1995).

Grafo 4. Grafo conectado

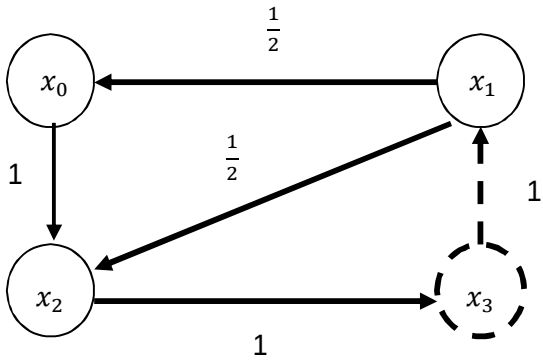


Como se puede observar, existen caminos dirigidos entre todos los nodos, pero comenzando por el nodo x_3 es imposible llegar a cualquiera de los nodos restantes.

Se dice 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.* (Beineke, Wilson, Cameron, 2005)

En un grafo fuertemente conectado partiendo desde cualquier nodo se puede llegar a cualquier otro nodo.

Grafo 5. Grafo fuertemente conectado



Como puede observarse, a diferencia de lo que sucedía en el grafo anterior, partiendo desde cualquier nodo es posible llegar a cualquier otro siguiendo el sentido de las flechas.

Esto tiene una implicancia muy importante para la matriz de adyacencia asociada a un grafo fuertemente conectado, ya que la sumatoria de los elementos de toda fila de la matriz es igual a uno. En símbolos:

$$\sum_{i=1}^{n} a_{x_j x_i} = 1 \qquad \forall\, j \;/\; j \in [1, n]$$

2. CADENAS DE MARKOV

En esta sección se exponen algunos conceptos esenciales sobre cadenas de Markov considerados necesarios para el desarrollo del presente trabajo. Se comenzará definiendo dos conceptos esenciales para luego tratar brevemente las Cadenas de Markov propiamente dichas.

En primer lugar, se denomina vector de probabilidad a aquel vector fila $p = (p_1, p_2, \dots, p_n)$ cuyos componentes son no negativos y la sumatoria de las mismas es igual a uno. Es decir, aquel vector que cumple con las siguientes condiciones:

i) $p_j \geq 0 \ \forall \ j$

ii) $\sum_{j=1}^{n} p_j = 1$ (Ley de Cierre)

Luego, una matriz estocástica o de probabilidad es aquella matriz cuadrada $P = [p_{ij}]_{nxn}$ en la cual cada una de sus filas es un vector de probabilidad.

El segundo concepto es el de proceso estocástico. Se entiende con este nombre a aquel fenómeno aleatorio que se desarrolla en el tiempo obedeciendo a leyes probabilísticas. Se puede pensar también un proceso estocástico como una sucesión de variables aleatorias que brindan una descripción de un determinado fenómeno a través del tiempo. En términos más formales, podemos escribirlo así:

$$\{X(t); t \in T\} \quad X(t): estado\ del\ proceso\ en\ t \quad T: conjunto\ temporal$$

Habiendo realizado estas definiciones, se puede definir un proceso o cadena de Markov como un proceso estocástico donde el valor presente de una variable aleatoria es relevante para predecir el valor futuro de dicha variable. Esto nos indica que la probabilidad de realización de determinado suceso solo depende de lo ocurrido en el periodo inmediatamente anterior. (Bharucha-Reid, 1960).

Desde el punto de vista probabilístico, se pueden pensar estos procesos entendiendo que en ellos la probabilidad de ocurrencia de un suceso en el período siguiente, dados los sucesos que acontecieron en los períodos anteriores (t, t-1,…), es igual a la probabilidad de ocurrencia de un suceso en el período t+1 dado lo sucedido en el período inmediatamente anterior. Es decir, la realización de un suceso en el período t+1 es independiente de lo acontecido en períodos anteriores a *t* (*t-1,..., t-n)*.

Su notación estadística sería entonces: $P(X_{t+1}|X_t; X_{t-1}; X_{t-2}; \dots; -n) = P(X_{t+1}|X_t)$

Ahora, esta probabilidad de que un proceso que se encuentra en el estado *i* (i = 1, 2,3,...) en el momento *t-1* pase al estado *j* (j = 1, 2,3,...) en el momento *t* (probabilidad de transición simple), puede ser expresada de la siguiente manera:

$\Pr[(T_t=j)/(T_{t-1} = i)] = P_{ij}(t-1; t), (i, j = 1,2,3, \dots; t = 0,1,2, \dots)$

Si se ordenan estas probabilidades según sus estados de partida y llegada se puede definir una matriz estocástica llamada "matriz de transición" P_t:

$$\mathrm{P_t} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

Donde p_{ij} es la probabilidad de pasar del estado i al j y en la cual, por ser $\mathrm{P_t}$ una matriz estocástica, se cumple que $\sum_{i=1}^{n} p_{ji} = 1 \quad \forall j$ (Bharucha-Reid, 1960)

Para calcular cómo se comportará una variable en un periodo t cualquiera se utiliza la fórmula $S(t) = S(0).P^t$ donde P es la matriz de transición, S(0) es el vector de estado inicial y S(t) el vector de estado en el período t. Notar que tanto S(0) como S(t) son vectores filas con tantos componentes como cantidad de estados tenga la variable que se está analizando.

Existe un vector de probabilidad, denominado vector punto fijo *v*, en el cual se verifica que cualquier transición de acuerdo con la matriz P no tiene efecto sobre sus probabilidades. En otros términos, el vector punto fijo satisface la siguiente igualdad:

$$v.P = v$$

El vector será el resultado de la reiteración del proceso de cálculo de comportamiento de las variables definido al comienzo de esta sección. En otros términos, cuando t tienda a infinito, la fórmula $S(t) = S(0).P^t$ convergerá al vector punto fijo. Para evitar la realización de cálculos excesivos, existen dos métodos para calcular el vector: mediante la resolución por sistema de ecuaciones o por el método de la adjunta. En el anexo se puede encontrar un ejemplo desarrollado del segundo método. (Bernardello, A., Bianco, M. J.: 2010)

3. LA RED Y LA BÚSQUEDA

3.1 La red de páginas web como un grafo

La aplicación de los conceptos expuestos en la primera sección a la red de páginas web es directa.

En este caso los nodos son las páginas web y los vínculos entre ellas son los hipervínculos que dirigen al usuario desde una página hacia otras.

Por las características propias de las páginas web necesariamente el vínculo entre páginas tiene una orientación definida ya que se parte desde la página *x* y se dirige hacia la página *y*. En consecuencia el grafo es orientado.

Aún más, al considerar la totalidad de vínculos existentes en una página hacia cualquier otro conjunto de páginas, fácilmente se puede asignar la ponderación correspondiente a cada vínculo. La forma más común, cuya justificación se realizará en la subsección siguiente, es realizarlo de la manera equitativa, aunque nada impide que existan casos en los cuales existan otros factores que incidan en las ponderaciones, y por lo tanto éstas no sean iguales.

Por último, no se puede aseverar que la representación de cualquier parte de la red sea un grafo conectado o fuertemente conectado, ya que depende del subconjunto de páginas que se selecciona. Sin embargo, la red

en su totalidad es de una complejidad tal que el grafo que la representa ni siquiera es un grafo conectado.

3.2 Recorriendo la red de páginas web

Habiendo presentado la utilidad de la teoría de grafos para representar la red de páginas web, ahora es preciso indagar en el accionar de la persona que va recorriendo este conjunto de páginas para encontrar lo que desea.

Al empezar a buscar algo en las páginas web, necesariamente se debe comenzar por alguna página. En caso de no encontrar lo deseado allí uno puede probar suerte eligiendo una nueva página al azar o seleccionar algunos de los hipervínculos presentes en la actual página. La mayoría de de las personas apostaría a la segunda opción ya que existen mayores chances de encontrar así lo que se desea.

Cada persona va a seleccionar el vínculo que considere más relevante para el objeto de su búsqueda y procederá de esta manera hasta encontrar lo que desea. Pero para ir introduciendo a los motores de búsqueda consideremos el caso en el cual el individuo elige los vínculos de manera aleatoria. Así se puede modelar la búsqueda web como un *random walk.*

En este caso, al estar en una determinada página la elección de cualquiera de los vínculos depende de la cantidad de vínculos existentes. Al modelar la búsqueda como un *random walk* la probabilidad de elegir cualquiera de los vínculos es igual para cada uno de ellos, dada la página en la que se encuentran. En símbolos: Sea X la variable que indica en qué página web se encuentra el usuario, i la página en la que se encuentra el usuario, J el número total de páginas con vínculos en la página i y el subíndice t indica los distintas "etapas" de la búsqueda, entonces:

$$P(X_{t+1} = j | X_t = i) = \frac{1}{J} \qquad \forall j \in J$$

En la etapa *t* se encuentra en la página *i* y en *t+1* en cualquiera de las *j* páginas con la misma probabilidad de ocurrencia.

Consideramos importante remarcar que se reconoce que existen algunas cuestiones subjetivas en la selección de los hipervínculos por parte del usuario, pero a los efectos del trabajo éstas serán obviadas al no ser lo suficientemente relevantes.

3.3 Recorriendo el grafo

Considerando a la red de páginas web como un grafo y a la búsqueda en ella como un fenómeno aleatorio, lo que en una sección aparecía como la ponderación del vínculo es lo que en la siguiente aparecía como la probabilidad de transición de una página a la próxima. Queda así justificado el motivo por el cual se trabaja con ponderaciones equitativas para los vínculos.

En consecuencia, la matriz de adyacencia de un grafo de páginas web es la matriz de transición asociada al proceso de búsqueda en estas páginas web. En este punto cobra importancia hacer la distinción entre los grafos conectados y fuertemente conectados. En los primeros, si la página presenta al menos un vínculo de salida, se sabe que la sumatoria de los elementos de la fila que representa los vínculos de salida de dicho nodo es igual a uno. Pero nada nos asegura que esto suceda, ya que puede existir un nodo que no tenga ningún vínculo de salida, lo cual se ve reflejado en el hecho de que la fila en la matriz de adyacencia se compone únicamente de ceros.

Debido a las características de los grafos fuertemente conectados en todas las filas de la matriz de adyacencia se verifica que la sumatoria de sus elementos es igual a uno. Como consecuencia, se puede asegurar que:

La matriz de adyacencia asociada a un grafo fuertemente conectado cumple con las características de una matriz estocástica.

Entonces, la matriz de adyacencia de un grafo fuertemente conectado es la matriz de estocástica asociada al proceso de búsqueda en las páginas web, considerando a éste como un fenómeno aleatorio.

Este resultado es central para las explicaciones siguientes y, a menos que se indique lo contrario, se trabaja con matrices que representan grafos fuertemente conectados.

4. ALGORITMO PAGE RANK

Para realizar búsquedas en la web se utilizan básicamente dos sistemas: los Test Based Ranking Systems y Page Rank. A continuación explicaremos brevemente el primero para luego abordar en profundidad el sistema Page Rank.

i) Test Based Ranking Systems: son los sistemas de ranking basados en texto y fueron utilizados principalmente en los años 90. Este tipo de motor selecciona aquellas páginas en las que aparezca más veces la palabra buscada y las ordena de forma decreciente. Presenta el inconveniente de que tal vez la página web en la que aparezca más veces el término buscado no sea relevante para la búsqueda o no aporte información significativa.

Se hace necesario un motor de búsqueda que de alguna manera filtre las páginas irrelevantes o no relacionadas para con la búsqueda y que solamente devuelva aquellas páginas que resultaran útiles y aportaran información.

ii) Page Rank: es uno de los algoritmos de búsqueda más populares e influyentes de la actualidad. Fue inventado por Larry Page y Sergey Brin y es el sello distintivo de Google desde 1998.

El algoritmo Page Rank se desprende de la idea de que se puede juzgar la importancia de una página web mirando las páginas que contienen un vínculo hacia la misma. Si una página A contiene un vínculo hacia otra página B se interpreta que la pagina A considera que el contenido de B es relevante para la temática abordada en A. Si existen muchas páginas con links hacia B se considera que es de común acuerdo que la pagina B es

importante. Por otro lado, si la página B tiene solamente un backlink[1] pero este proviene de una página C con *autoridad* (como www.bbc.com o www.cnn.com) decimos que C transfiere su autoridad a B, es decir, indica que B es importante. Utilizando estos conceptos de importancia y autoridad el algoritmo page Rank asigna un rango a cada página basándose en las páginas que dirigen a ellas.

La gran extensión de la web lleva a utilizar algoritmos cuya complejidad excede el alcance de este trabajo. Sin embargo, se puede alcanzar una comprensión de los rudimentos del funcionamiento de estos algoritmos partiendo de una versión acotada de la web.

Si se parte de una representación de la web mediante una red de grafos fuertemente conectados se podrá representar la web con una matriz estocástica.

Se considera que inicialmente la probabilidad de acceder a una página está igualmente distribuida entre todas las páginas de la web. Sin embargo, al comenzar el proceso de navegación o de búsqueda, la probabilidad de llegar a cada una de las páginas se verá modificada por la composición de la matriz de transición previamente elaborada. En otros términos, como se ha establecido, cada *backlink* incrementa la importancia de una página, por lo que se puede actualizar el rango de cada página adicionando a su valor corriente (el valor actualizado hasta ese momento) la importancia transferida por los vínculos que dirigen hacia ella.

Si por cada pasaje de una página a la otra se actualiza el vector inicial según la fórmula $S(t) = S(0).P^t$, luego de cierto número de iteraciones se alcanzará un vector de importancia de equilibrio que se denominará vector Page Rank de las páginas. El algoritmo interpretará este vector de

[1] *Backlink* o vínculos externos de respaldo, son los enlaces que recibe una página web desde otras páginas web. El número de *backlinks* es la cantidad de páginas que la enlazan a través de un vínculo.

importancia de equilibrio de forma tal que aquella página que obtenga un mayor coeficiente (es decir, una importancia más alta) será la que obtenga un mejor puesto en el ranking y aparecerá entre los primeros resultados del buscador.

Es importante aclarar que tanto el razonamiento como método presentados solo son válidos para situaciones "normales", donde todas las páginas se entrelazan entre sí y pueden ser representadas por grafos fuertemente conectados. Para aquellos casos en los que esto no suceda, el algoritmo tiene fórmulas de corrección que permiten su correcto funcionamiento. Dos inconvenientes comunes son:

- Páginas sin links de salida (páginas colgantes), correspondientes a grafos conectados.
- Presencia de componentes desligados, es decir, grupos de páginas que no estén relacionados entre sí y que sin embargo sean relevantes uno para con el otro.

En resumen, se puede interpretar el proceso de búsqueda web como un proceso o cadenas de Markov, es decir, un proceso estocástico en el cual cada cambio de página es independiente de los sitios visitados anteriormente. Como fue expuesto en la sección anterior, una web fuertemente conectada puede ser representada por una matriz estocástica y la misma puede ser interpretada como la matriz de transición correspondiente al pasaje de una página a otra. Consecuentemente, el vector de importancia de equilibrio puede ser interpretado como el vector punto fijo correspondiente a la matriz de transición.

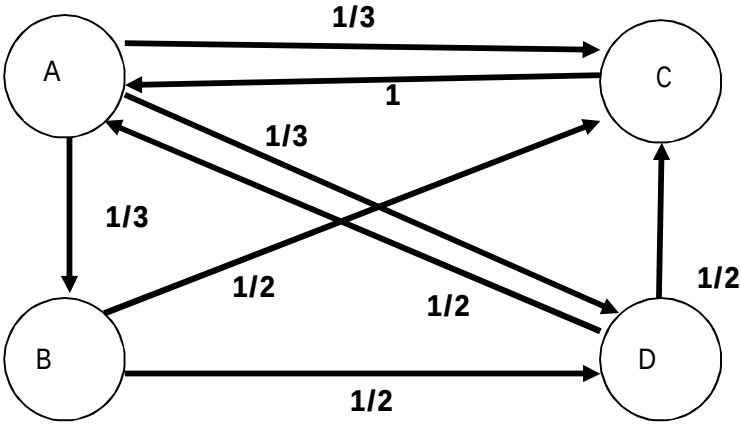
5. EJEMPLO NUMÉRICO

A continuación se expondrá un ejemplo sencillo para comprender los rudimentos de cómo el algoritmo ranquea las páginas. Se utilizarán grafos para representar la red, tal y como fue expuesto en la sección II.

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

Se adiciona la ponderación que le corresponde a cada página suponiendo que la importancia está distribuida de manera equitativa entre las páginas.

Grafo 6



Se puede observar que el grafo se encuentra fuertemente conectado. Luego, en base al mismo, se construye su matriz de adyacencia asociada. Recordar que por tratarse de un grafo fuertemente conectado dicha matriz es estocástica.

$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Se supone que inicialmente la importancia está distribuida uniformemente entre las cuatro páginas, se define: $v = \left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right)$

Siendo *v* el vector de ranking inicial.

Se multiplica la matriz P por el vector v para actualizar la importancia.

Se obtiene entonces un nuevo vector de importancia $v_1 = vA$. Se repite el proceso llegando a un nuevo vector de importancia actualizado con la forma $v_2 = (vA)A = vA^2$. Reiterando infinitamente este proceso, se converge al vector de importancia de equilibrio.

Lo que se busca con estos procesos de actualización es recrear el camino que recorrería un usuario de la web en su búsqueda, es decir, su desplazamiento entre las páginas.

Para el ejemplo analizado:

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^8 = (0,38 0,12 0,29 0,19)

Este es también un vector muy próximo en sus valores al vector de importancia de equilibrio (v*)

Si se interpreta el proceso de búsqueda web como un proceso de cadenas de Markov, considerando a P como la matriz de transición correspondiente al modelo de cuatro páginas web, se procede a calcular el vector punto fijo por cualquiera de los métodos conocidos y se observa que éste coincidirá con la aproximación al vector de importancia de equilibrio alcanzado anteriormente.

Vector punto fijo: v*= (0,38 0,12 0,29 0,19)

Se comprueba entonces que el vector de importancia de equilibrio es el vector punto fijo asociado a la matriz de transición correspondiente con la red de páginas. Esto facilita los cálculos a la hora de buscar *v** y así ordenar las páginas según su importancia.

La interpretación del mismo nos indica 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.

6. CONCLUSIONES

Los actuales algoritmos de búsqueda suponen una enorme ventaja en comparación con los viejos algoritmos para lograr un mejor ordenamiento de las páginas web al momento de realizar una determinada búsqueda. La obtención del vector de importancia de equilibrio a partir de las relaciones existentes entre las páginas permite ordenar las páginas web de una manera inmensamente más útil que aquella que se puede lograr mediante un algoritmo de búsqueda basado en texto.

Se mostró la gran utilidad de los grafos para representar la red de páginas web y de las cadenas de Markov para modelar la búsqueda en la red como un proceso aleatorio. Así se ha logrado abordar, mediante conceptos

matemáticos básicos, una parte de la compleja temática de la búsqueda web.

Otro aporte relevante es que se muestra una aplicación de las Cadenas de Markov en la cual el fenómeno aleatorio abordado no se desarrolla en el tiempo, como es lo habitual con este tipo de procesos. En este caso los algoritmos de búsqueda se limitan a simular infinitas búsquedas en un instante para llegar al vector de importancia de equilibrio en un brevísimo lapso.

REFERENCIAS BIBLIOGRÁFICAS

Beineke L. W., Wilson R. J., Cameron P. J. (2005). *Topics in Algebraic Linear Theory*. Nueva York: Cambridge University Press. Recuperado de http://catdir.loc.gov

Bernardello, A., Bianco, M. J., Casparri, M. T., Fronti, J. G., Olivera de Marzana, O. (2010). *Matemática para Economistas utilizando Microsoft Excel y MATLAB*. Buenos Aires, Omicron System.

Biggs, N. (1993). *Algebraic Graph Theory.* Cambridge, Cambridge University Press. Recuperado de https://books.google.es

Bollobás, B. (1998). *Modern Graph Theory.* Nueva York, Springer. Recuperado de https://books.google.es

Cox, D.R., Miller H.D. (1970). *The Theory Stochastic Processes*. Methuen.

Bharucha-Reid, F A.T. (1960). *Elements of The Theory of Markov Processes And Their Applications*. McGraw Hill Series in Probability and Statistics.

ANEXO
MÉTODO DE LA ADJUNTA PARA EL CÁLCULO DEL VECTOR PUNTO FIJO

Este método de resolución se basa en la propiedad de las matrices estocásticas enunciada anteriormente, según la cual toda matriz estocástica cuenta con un autovalor igual a uno.

La forma de proceder es la siguiente:

I) Se resta a la de transición P la matriz identidad de manera que queda una nueva matriz $(P - I)$.

II) Se elige una de las columnas de $(P - I)$ y se confecciona un nuevo vector denominado W a partir de los menores adjuntos de los elementos de la columna seleccionada; es decir, el primer elemento w_1 es el menor adjunto del primer elemento de la columna y así sucesivamente. Tener en cuenta que cada uno de los elementos es un número real. Un ejemplo genérico sería:

$$W = (w_1 \quad \dots \quad w_n)$$

III) El vector W no cumple necesariamente con la Ley de cierre ($\sum_{i=1}^{n} w_i$ puede que no sea igual a 1), con lo cual el mismo no es un vector de probabilidad y en consecuencia no puede ser un vector estacionario o punto fijo.

Sin embargo todo vector que cumpla con la condición de que todos sus elementos son no negativos y cuenten con al menos un elemento positivo, tienen un vector único de probabilidad. La forma de llegar a dicho vector es dividir cada elemento por la sumatoria $\sum_{i=1}^{n} w_i$.

Luego el vector punto fijo será:

$$F = \left(\frac{w_1}{\sum_{i=1}^{n} w_i} \quad \dots \quad \frac{w_n}{\sum_{i=1}^{n} w_i} \right)$$