

Cómo ordena el buscador Google sus resultados.

Roberto Markarian-Nelson Möller
IMERL- Facultad de Ingeniería
Universidad de la Republica-URUGUAY

RESUMEN

Cuando deseamos encontrar alguna información en Internet, solemos usar “buscadores” de fácil acceso que tienen diversos nombres comerciales: Yahoo, Google, etc. Generalmente sólo miramos los primeros resultados que se nos presentan. Por ello es importante saber cómo se ordenan las variadas y muchas páginas que tienen algo en común con los temas o nombres consultados. En este trabajo explicamos un procedimiento que asocia a cada página de la Red un número que cuantifica su “relevancia” (o “importancia”) y permite ordenar los resultados de la búsqueda. Los enlaces de las páginas de la Red, determinan vectores cuyas propiedades permiten utilizar sus componentes como una medida de la “importancia”.

Este método popularizó el uso del buscador Google. Nuestro objetivo es ilustrar cómo son utilizadas ciertas nociones de matemática en estos cálculos¹.

1. INTRODUCCION

La siguiente cita, extraída del periódico Le Monde Diplomatique [La], ilustra en términos generales las ideas que profundizaremos.

“A mediados de los '90, frente al creciente flujo de información, dos estudiantes de computación de la Universidad estadounidense de Stanford, Sergey Brin y Larry Page, intuyeron algo: un motor de búsqueda que se basara en el estudio matemático de las relaciones entre los diferentes sitios daría mucho mejor resultado que las técnicas rudimentarias que se empleaban entonces. Convencidos de que las páginas más *pertinentes* son las más frecuentemente citadas (las que los otros sitios ponen como referencia en su lista de enlaces en hipertexto) deciden hacer del tema su proyecto de estudios, estableciendo así las bases de un motor más *matemático*, al que bautizaron Google en el momento de crear su empresa, en setiembre de 1998.”

Al buscar material en Internet planteamos dos problemas simultáneamente:

- ¿Qué página tiene elementos relacionados con lo que buscamos?
- ¿Cómo se presenta (ordena) todo lo hallado?

En este trabajo, analizaremos uno de los elementos que utiliza el buscador Google para ordenar las páginas relacionadas con nuestra búsqueda. Para ellos se utilizan elementos de Algebra Lineal; una rama de la matemática que, generalmente se estudia en los primeros cursos universitarios. Hallar que elementos de la Red se relacionan con nuestra búsqueda es un problema que también se resuelve utilizando elementos del Algebra lineal, que no trataremos en este trabajo, ver [BDJ].

Los resultados de nuestras búsquedas suelen ser muchas páginas de direcciones relacionadas con el tema; pero, pocas veces miramos más allá de las primeras. Por ello es muy útil que estas primeras sean las más relevantes. Es allí donde interviene uno de los principales elementos introducidos por el Google en 1998, el PageRank [BP]:

¹ Algunos de los aspectos matemáticos que trataremos no son elementales, por lo que sugerimos adaptarse a las siguientes "instrucciones": El artículo puede ser leído, por lo menos, de tres maneras distintas: los que quieran evitar toda la matemática, deberán leer la Introducción y las Secciones 2 y 5; los que quieran evitarse muchas complicaciones matemáticas deberán dejar de lado las secciones 3 y 6 en adelante; quienes no quieran entrar en alguna matemática de nivel universitario, hasta la sección 5.

“Para medir la importancia relativa de las páginas web nosotros proponemos PageRank, un método para calcular un ordenamiento (*ranking* en inglés) para toda página, basado en el gráfico de la Red.”

En ese método, continúa el artículo citado de Le Monde,

“ la importancia de las páginas web es reevaluada permanentemente en función de la cantidad de menciones de que son objeto en los diferentes sitios. Por lo tanto, los sitios aislados, que no figuran en las listas de enlaces hipertextuales, resultan poco visibles, sin “legitimidad”. En cambio los sitios muy citados se convierten para Google en sitios de referencia. Ese original algoritmo ofrece resultados impresionantes.”

En los años 50, ya se había observado la posibilidad de utilizar este tipo de cálculos para obtener ordenamientos. Su vigencia actual se debe a su aplicación a Internet y a los desarrollos en el cálculo con objetos matemáticos de tamaño muy grande. Resulta interesante, en este sentido, tener en cuenta los montos destinados por el organismo promotor de la ciencia del gobierno de los Estados Unidos de América (NSF, National Science Foundation) a

los grupos de la Universidad de Stanford que trabajan en estos problemas: en los últimos tres años han recibido más de tres millones de dólares.

2. ALGO DE HISTORIA

La Red crece y se modifica en forma vertiginosa. Hagamos un poco de historia para situar el contexto de invención e implementación del procedimiento utilizado por Google.

En 1996-98 comenzaba a notarse la dificultad para hallar material en internet debido a su rápido crecimiento. En ese momento “buscadores”, también llamados “motores de búsqueda”, como Altavista, Lycos, Yahoo, etc., ya tenían gran relevancia.

En principio, todo buscador comprende por lo menos tres elementos principales:

- un robot de indexación (también conocido como *araña*, *spider* o *web crawler*),
- una base de datos, y
- una interface de consulta de la base de datos.

Normalmente, los usuarios interactúan con la interface de consulta y, a través de ella, consultan la base de datos. El robot de indexación “navega” en la Red colectando toda la información que este pueda procesar y almacenándola para su posterior procesamiento y consulta.

A mediados de los 90, los buscadores desarrollaban tecnologías para restringir la búsqueda. Estas restricciones empleaban argumentos lógicos que no eran de manejo sencillo. Yahoo, que también surgió en Stanford, hacía “manualmente” el trabajo de ordenar de acuerdo a ciertos criterios “objetivos” las bases de datos disponibles. Dichas bases de datos tenían un tamaño considerable, por lo que ya estaba muy popularizado el uso de buscadores. Los que funcionaban bien eran un gran negocio: Yahoo se vendió en una abultada cifra por esa fecha. Los algoritmos de búsqueda recibían un gran impulso y a pesar de ello no se simplificaban los procedimientos de ordenación.

En ese contexto, y en pleno boom de las compañías puntocom, fue que comenzó en la Universidad de Stanford la historia de Google.

Sergey Brin y Lawrence Page presentaron un trabajo de postgrado donde se definía la “importancia” de una página web. Para ello consideraban los enlaces que la misma recibía. Su buscador hace una lista de respuestas a nuestra búsqueda en un “orden de relevancia” decreciente. Esta fue la mejora en su interface de consulta que popularizó su uso. Hemos puesto el comillado porque se señalan deficiencias y críticas al modo cómo se hace la cuantificación (de la “relevancia”). Algunas de éstas serán comentadas más adelante.

3. COMO ORDENAR LAS PÁGINAS DE LA RED.

Estando en una página web T tenemos dos números importantes:

- cantidad de vínculos entrantes = cantidad de páginas que tienen un vínculo hacia la página T ;
- cantidad de vínculos salientes.

Las páginas Web varían mucho en el número de vínculos entrantes que poseen. Generalmente, las páginas que son apuntadas desde muchas páginas son más importantes que las páginas a las cuales sólo se llega desde unas pocas. Pero, hay muchos casos en los que sólo el contar el número de vínculos entrantes no se corresponde con el sentido usual de la importancia de una página Web.

Como escribían Brin y Page [BP]:

“Por ejemplo, si una página tiene un vínculo de la página principal de Yahoo, éste puede ser un solo vínculo pero uno muy importante. Dicha página debería estar mejor clasificada que otras páginas con muchos vínculos pero de lugares desconocidos”.

Por tanto, una página tiene una clasificación alta si la suma de las clasificaciones de sus vínculos entrantes es alto. Esto cubre ambos casos:

muchos vínculos entrantes o unos pocos con alta clasificación.

El algoritmo original del PageRank fue descrito en varios trabajos por Brin y Page [BP]. Posteriormente, presentaron una versión mejorada, que es la que expondremos.

El propósito es cuantificar la probabilidad de que un usuario (aleatorio) llegue a la página T utilizando la Red. Se define el PageRank de una página T por:

$$PR(T) = \frac{1-d}{N} + d \left(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

donde:

- N = número total de páginas Web desde las que salen vínculos.
- n = número total de páginas Web desde las que salen vínculos a la página T .
- $PR(T_i)$ = PageRank de las páginas T_i que tienen un vínculo hacia la página T .
- $C(T_i)$ = número de vínculos salientes de la página T_i .
- d es un factor de amortiguación que puede ser tomado entre 0 y 1.

A efectos de entender mejor esta fórmula, le recomendamos hacer el ejercicio de sumar todos esos números. Todos, significa sumar los $PR(T)$ sobre todas las páginas Web. Si esa suma da h , obtendrá la siguiente fórmula :

$$h = (1-d) + d h,$$

lo que implica que $h=1$. Por ello se dice que $PR(T)$ es una distribución de probabilidad (indexada por el parámetro d). Esta “normalización” (suma=1) facilita la utilización de resultados generales que no dependen del tamaño del sistema (el número total de páginas).

Analizando con cuidado dicha fórmula se observarán las siguientes características del PageRank:

- se define para cada página y depende de los PageRanks de las páginas que tienen un vínculo hacia ella;
- los sitios que enlazan a la página T no influyen uniformemente pues depende del número de vínculos salientes que ellas posean: a más vínculos salientes de una página menos beneficiará el PageRank de las páginas a las que se una;

- un nuevo vínculo a una página siempre aumenta su valor;
- la definición es recursiva: la clasificación de una página depende de todas las otras que tienen vínculos hacia ella, por ello la clasificación de cada página depende de todos los sitios de la Red.

Brin y Page en sus explicaciones dan una justificación sencilla para el algoritmo. El PageRank modela el comportamiento de un usuario que estando en una página puede:

- elegir al azar entre los vínculos contenidos en la página actual, o
- saltar al azar a cualquier página de la Red ingresando la dirección;

todo ello sin considerar el contenido de los mismos (esto ha suscitado comentarios y modelos alternativos que no podemos comentar en este artículo). Se supone que sigue un enlace de la página en que está con probabilidad d , o salta a cualquier página de la red con probabilidad $1-d$.

Parece razonable suponer que d es mayor que un medio, o sea que, estando en una página, se tiende a usar más los vínculos que allí están, que hacer una nueva elección al azar.

La única excepción son las páginas hacia las que no va ningún vínculo, a las cuales en este modelo, por estar aisladas, sólo se llega al azar. No caben dudas que a ellas se puede llegar buscándolas explícitamente, pero para usar este procedimiento -que es el mejor procedimiento de búsqueda- no se necesitan “buscadores”. El PageRank de estas páginas es $(1-d)/N$.

La definición del PageRank establece un procedimiento para determinar una probabilidad de que un usuario aleatorio llegue a la página Web T . Este usuario visita una página web con una probabilidad proporcional al PageRank de la página. La probabilidad de elegir un vínculo, se distribuye uniformemente entre los vínculos que tiene para elegir en la página.

Una ventaja de esta definición es que posibilita utilizar un algoritmo iterativo que aproxima los valores de PageRank. O sea, a cada página se le asigna un valor inicial y se realizan iteraciones que modifican sucesivamente estos valores iniciales. Esto es, a partir de distribuciones iniciales prefijadas, se repite un mismo procedimiento para obtener nuevos valores para cada página, y así sucesivamente. Este es un punto importante a la hora de implementar el mecanismo, pues en términos computacionales es más sencillo calcular iterativamente el valor deseado que mediante otros procedimientos.

Otra ventaja es que este método establece un ordenamiento predeterminado, no hay que realizar el cálculo cada vez que alguien realiza una búsqueda. También relega el papel de los enlaces salientes lo que lo hace menos sensitivo al spamming.

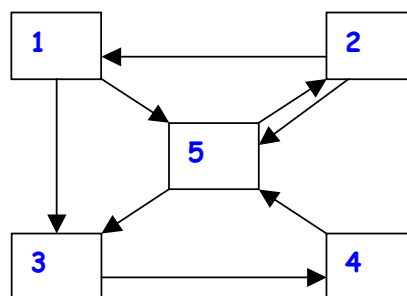
Algunas preguntas surgen naturalmente;

- ¿Por qué este procedimiento funciona?
- ¿Será que este procedimiento lleva a dar a cada página un valor único, su PageRank?

Las respuestas afirmativas, en general, incluyen el uso de una versión del teorema de Perron-Frobenius que se enuncia en el Apéndice y se trata con detalle en [MM].

4. UN EJEMPLO

Comenzamos con una versión simplificada del problema dada por el siguiente diagrama.



Tenemos 5 páginas web e indicamos con una flecha los vínculos. Por ejemplo, de la página 1 salen dos vínculos a las 3 y 5, y entra un vínculo de la página 2.

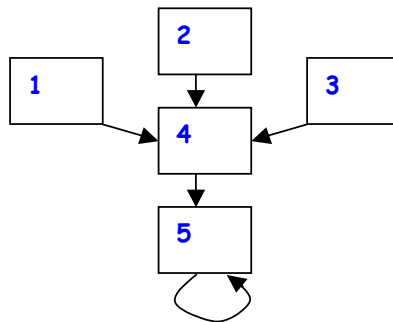
Veamos las fórmulas de PageRank para este caso . Llamamos $PR(i)$ al PageRank de la página i :

$$\begin{aligned} PR(1) &= \frac{1-d}{5} + d \left(\frac{PR(2)}{2} \right) \\ PR(2) &= \frac{1-d}{5} + d \left(\frac{PR(5)}{2} \right) \\ PR(3) &= \frac{1-d}{5} + d \left(\frac{PR(1)}{2} + \frac{PR(5)}{2} \right) \\ PR(4) &= \frac{1-d}{5} + d \left(\frac{PR(3)}{2} \right) \\ PR(5) &= \frac{1-d}{5} + d \left(\frac{PR(1)}{2} + \frac{PR(2)}{2} + PR(4) \right) \end{aligned}$$

Haciendo un cálculo que explicaremos un poco más en el Apéndice, con $d=0.85$, llegamos a los siguientes valores,

$$\begin{aligned} PR(1) &= 0.1003570039 \\ PR(2) &= 0.1655458921 \\ PR(3) &= 0.2081976187 \\ PR(4) &= 0.2069679755 \\ PR(5) &= 0.3189315099 \end{aligned}$$

Observe que la página 5 es la que tiene mejor clasificación. Si se realiza el cálculo con un esquema del tipo que sigue, se verá que nuevamente la página 5 será la más relevante.



¿Qué sucede si la página 5 no enlaza consigo misma?

En ese caso la página 5 representa lo que se conoce como *enlace colgado*. Se observará que en la expresión de $PR(4)$ aparece una división por cero pues se deberá poner $PR(5)$, vínculo saliente de cinco, que es cero. En virtud de que no se puede dividir por cero, el cálculo debe modificarse, se realiza el cálculo sin tomar en cuentas las páginas que corresponden al enlace colgado y luego se hallan la calificación de los “colgados”.

5. GOOGLE EN SERIO

Se considera el conjunto W de páginas que se pueden alcanzar apartir de una página en Google. Sea N el número de páginas en W , este número varía con el tiempo. La página de Google indica la cantidad de páginas que posee indexadas; actualmente N es superior a 4200 millones.

Implementar el algoritmo para aproximar el PageRank de todas las páginas de forma recursiva no es algo menor, estamos hablando de manejar una matriz que tiene un tamaño de varios millones. En el Apéndice se indicará porqué funciona esta implementación que asigna una calificación no nula única a cada página.

En teoría toda página posee un PageRank positivo, pero, en el ordenamiento real se introducen como penalización una calificación nula llamada **PR 0**.

Las páginas que aparecen en los primeros lugares de un listado de Google, generalmente, tienen mayor número de visitas que aquellas que aparecen relegadas. En esto radica el interés de los responsables (*webmaster*) de las páginas comerciales u otras por hacer aparecer sus sitios en los primeros lugares, intentando aumentar las calificaciones de sus páginas a través del manipuleo de su enlaces. A su vez, los administradores de Google quieren evitar trampas de este tipo, por lo que se intenta detectar y penalizar tales intentos.

Públicamente se desconoce la forma en que se realiza, puesto que, diversos elementos que hacen funcionar su buscador son secretos comerciales. En la Red, existe material que especula acerca de la implementación de esta penalización [EF].

Debido a la naturaleza del orden que establece el PageRank,

una búsqueda no lleva hacia la referencia “principal” sobre el tema sino hacia la acepción más ampliamente citada.
--

En los términos utilizados en Le Monde [La]:

“En realidad, el poder de influencia de los diferentes actores depende sobre todo de su grado de apropiación de la Red: no alcanza con desarrollar un sitio, también hay que ser capaz de establecer vínculos con los otros sitios y obtener el reconocimiento de 'los que cuentan' en internet.”[...]

“Es sin duda en los temas políticos -sobre los cuales cohabitan en internet puntos de vista radicalmente diferentes donde Google pone de manifiesto sus límites: sus criterios matemáticos pueden privilegiar *de facto* ciertas opiniones y brindar una pertinencia indebida a textos que sólo representan la opinión de unos pocos.”

En este sentido, los “adelantados” de internet, los “gurus” del fenómeno esencialmente estadounidense de los *weblog*, llevan las de ganar por la densidad de lazos que pueden establecer y las bases de datos que manejan. Se han realizado experiencias exitosas que muestran las posibilidades de utilizar “artificialmente” esta característica para subir el PageRank de una página.

Uno de los casos más conocidos es el de “miserable failure”(falla miserable). Si se comienza (en mayo de 2004) una búsqueda con esas palabras se verá que en primer lugar aparece el sitio de la Casa Blanca que contiene la biografía de George Bush. Esto fue hecho a iniciativa de un usuario que conocía bien los procedimientos de calificación. Estos casos se conocen como “Google Bombing”.

En este momento, Google no sólo es el buscador más utilizado sino que, vende servicios a portales importantes: Yahoo, AOL, etc. Además, su sistema llamado de publicidad direccionada (cuando usted introduce palabras para buscar, junto con los resultados de su búsqueda Google presenta propaganda relacionada con lo buscado) es la que dirige mayor cantidad de gente hacia sitios comerciales². Se estima que, por venta de servicios y licencias de su tecnología de búsqueda tiene ganancias por 150 millones de dólares [Ec]. Un elemento no menor luego de la caída de las puntocom de marzo 2000.

La empresa Google Inc intenta mantener su preponderancia en el mercado de buscadores. La empresa Kaltix, también formada en Stanford en Junio 2003, fue adquirida por Google a fines de septiembre del 2003. Esta empresa desarrollaba tecnologías de búsqueda personalizadas y sensitivas al contexto que las aceleran y las hacen de más fácil utilización.

² Este hecho, así como una descripción de comportamientos que han sido punidos por Google Inc se pueden consultar en Wall Street Journal [WSJ].

Google, empresa, lanzó su cotización en el mercado electrónico Nasdaq (bajo el código GOOG), luego de un proceso³ que despertó muchas expectativas. El 17 de agosto, día previo al lanzamiento, debido a las críticas y a las malas condiciones del mercado, la empresa californiana anunció que rebajaba el rango de precios de venta a 85 y 95 dólares, si bien finalmente eligió el más barato. Posteriormente, en una situación que recuerda los mejores tiempos de la 'burbuja tecnológica', la cotización se elevó un 60 por ciento en solo unos minutos.

El desembarco en bolsa del buscador líder de Internet acaparó la atención de los inversores, porque es una de las principales colocaciones desde que estalló la llamada 'burbuja tecnológica'. Las acciones de la empresa californiana, marcaron su primer precio en 135,91 dólares, si bien después sufrió una gran volatilidad que le llevaron a moverse entre los 95 y los 140 dólares. Esta volatilidad, típica luego de la venta de acciones por parte de una empresa, puso en tela de juicio los argumentos de los “expertos”. Estos, en las semanas previas al lanzamiento, criticaron los altos precios sugeridos por Google para sus acciones entonces situados entre 108 y 135 dólares.

En esta colocación, la empresa ha distribuido entre los inversores un total de 19,6 millones de acciones. De esta cantidad, 5,5 millones corresponde a los principales accionistas de la empresa, que han visto recompensada su apuesta de los últimos años, con una cantidad multimillonaria. En concreto, los fundadores de la compañía, Larry Page y Sergey Brin, han recibido cada uno 40 millones de dólares, mientras que el consejero delegado, Eric Schmidt, ha obtenido 30 millones.

En general, la empresa ha recaudado con esta Oferta Pública de Venta (IPO, por sus siglas en inglés) un total de 1.670 millones de dólares. La fuerte subida de las acciones de Google ha hecho que la empresa californiana pase momentáneamente de tener un valor de 23.100 millones a un total de 36.960 millones, más que la tienda de Internet Amazon y la tecnológica Lucent Technologies juntos.

Se dice que Microsoft, también estaría por lanzar su propia tecnología de búsqueda.

En junio de 2002, la Comisión Federal de Comercio de los Estados Unidos estableció ciertas reglas recomendando que cualquier ordenamiento influido por criterios monetarios más que por criterios “imparciales” y “objetivos” debía ser claramente indicado para proteger los intereses de los consumidores. Aún así, la apariencia “objetiva” de estos mecanismos debe ser cuestionada.

Para terminar esta sección resulta interesante mencionar algunos datos sobre Google Inc:

- atiende 150 millones de búsquedas por día,
- en sus picos sirve 4000 consultas por segundo,
- utiliza 15000 computadoras tipo PC.

Google es uno de los pocos motores de búsqueda que recorre la Red frecuentemente para mantener actualizada su base de datos (por lo menos así lo ha hecho en los últimos dos años). Lleva, aproximadamente, una semana cubrir la Red y otra para calcular el PageRank. El ciclo de puesta al día de Google es de aproximadamente 30 días. Se ha advertido, que el PageRank vigente influye el recorrido mensual realizado por Google: páginas con mayor PageRank son recorridas más rápidamente y “con mayor profundidad” que otras con menor clasificación.

Este último punto, hace que se vea como discriminatoria la naturaleza del PageRank [La], [Bra]. Se llega a afirmar que, los nuevos sitios lanzados en el 2002 tienen mayor dificultad en conseguir tráfico que antes que Google tomara una posición dominante y que la estructura de enlaces de la Red han cambiado significativamente a partir del predominio del Google.

³ Ver, por ejemplo <http://www.google-ipo.com/>

6. APÉNDICE. UN POCO MÁS DE MATEMÁTICA

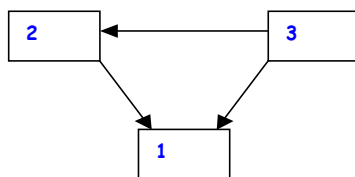
Aquí daremos algunos elementos que permitirán avanzar un poco en la comprensión de los problemas matemáticos del ordenamiento del Google. Somos conscientes que las dificultades matemáticas irán creciendo a medida que se avance en la lectura de este Apéndice.

MODELO SIMPLIFICADO.

Expresándolo de manera un tanto simplificado, lo que buscamos al calcular el PageRank es que *la importancia de cada página sea proporcional a la suma de las importancias de todos los sitios que enlazan con ella (tienen enlaces hacia ella).*

Esto se expresa, matemáticamente, de la siguiente manera. Llamemos α (la letra griega alfa) a la constante de proporcionalidad. Sea x_i el número que expresa la importancia de la página i -ésima.

Por ejemplo, si las páginas dos y tres están relacionadas con (enlazadas hacia) la página primera,



tendremos

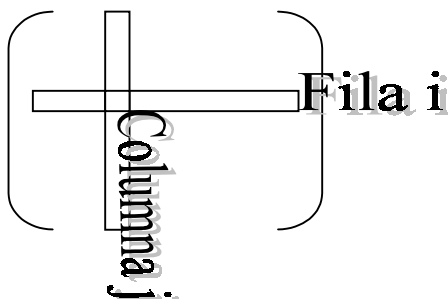
$$x_1 = \alpha(x_2 + x_3)$$

y así sucesivamente con todas las páginas. Quedan determinadas tantas ecuaciones como páginas estemos considerando y en cada igualdad el lado derecho será la suma de la importancia de todos los sitios que enlazan a la página del lado izquierda. Se tendrá lo que se llama un sistema de ecuaciones del tipo

$$\begin{aligned} x_1 &= \alpha(x_2 + x_3) \\ x_2 &= \alpha(x_3) \\ x_3 &= \alpha \cdots \\ &\vdots \end{aligned}$$

UNA PRESENTACIÓN RESUMIDA DE ESTAS FÓRMULAS.

Una matriz A es un ente matemático que se puede concebir con cierta facilidad. Se consideran $m \times n$ números ordenados en m filas y n columnas; cada uno de esos números se ubicará por la fila i y la columna j en que se encuentra



Así, un elemento genérico de la matriz se escribirá como a_{ij} . Observe que a_{11} es el primer elemento a la izquierda arriba de la matriz; y que a_{mn} es el de más abajo a la derecha. Por ejemplo, en la matriz siguiente que es de tamaño 5×5 , el elemento $a_{23} = 6$,

$$A = \begin{pmatrix} 0 & 1 & 3 & 3 & 4 \\ 0 & 5 & 6 & 7 & 8 \\ 1 & 9 & 10 & 0 & 11 \\ 20 & 30 & 21 & 0 & 40 \\ 13 & 14 & 10 & 15 & 40 \end{pmatrix}$$

Si llamamos n-uplas (vectores) a los conjuntos ordenados de n números, $x = (x_1, \dots, x_n)$, podemos definir el producto de la matriz A de n filas y n columnas por un tal vector x , de la siguiente manera

$$Ax = y,$$

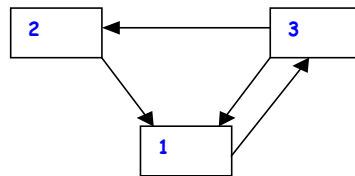
donde $y = (y_1, \dots, y_n)$ es una n-upla en que cada y_i se calcula así:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = \sum_{j=1}^n a_{ij}x_j$$

Por ejemplo, si A es la matriz anterior y $x = (1, 8, 0, -3, 2)$ tendremos

$$y_2 = 0 \times 1 + 5 \times 8 + 6 \times 0 + 7 \times (-3) + 8 \times 2 = 35.$$

Sea ahora A la matriz cuya entrada a_{ij} es 1 si el sitio j tiene un enlace hacia la página i y 0 en el caso contrario. Por ejemplo, si hubiera tres páginas tales que la segunda y la tercera se relacionan con la primera, la tercera con la segunda, la primera con la tercera, tendríamos



$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Si el número total de páginas de las que salen enlaces es n esta matriz nos permite reescribir las ecuaciones anteriores en la forma

$$x_i = \alpha \sum_{j=1}^n a_{ij}x_j \text{ que en forma matricial es } x = \alpha Ax.$$

Entonces, el problema de hallar vectores $x = (x_1, x_2, \dots, x_n)$ que satisfacen esa igualdad se transforma en hallar x tal que

$$Ax = \frac{1}{\alpha} x,$$

donde A es una matriz que toma en cuenta la estructura de vínculos o enlaces (*links* en inglés) de la Red. Obsérvese que x es un vector tal que Ax es un múltiplo de él; se dice que x es un *vector propio de la matriz A con valor propio* $1/\alpha$.

Así, el problema de hallar la importancia de una página, en este modelo simplificado, se transforma en un problema de vectores y valores propios de una matriz de vínculos entre las páginas de la Red.

Esta es una forma muy simplificada del funcionamiento real del Google. En realidad la matriz A tiene una forma más complicada no sólo por el tamaño (el número de páginas de la Red) sino por su estructura, que debe tener en cuenta otros factores, en particular la posibilidad de llegar al azar a una página cualquiera.

MATRIZ DEL PAGERANK.

Veamos cómo es la matriz de los PageRank del ejemplo de la sección 3, tomando $d=0.85$. Para uniformizar la nomenclatura llamamos $PR(i) = x_i$. Se deberá recordar que la suma de los PageRank es uno; o sea

$$\sum_{i=1}^5 x_i = 1,$$

por lo que cada sumando de la forma $(1-d)/5 = 0.15/5 = 0.03$ se puede escribir como $0.03 \sum_{i=1}^5 x_i$. Por ejemplo

$$PR(1) = x_1 = 0.03 \sum x_i + 0.425 x_2 = 0.03 x_1 + 0.455 x_2 + 0.03 x_3 + 0.03 x_4 + 0.03 x_5.$$

Entonces todas las fórmulas de las PageRank de la Sección 3 quedan escritas como $x = Px$, donde

$$P = 0.03 \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} + 0.85 \begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0.5 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.03 & 0.455 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.455 \\ 0.455 & 0.03 & 0.03 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.88 & 0.03 & 0.03 \\ 0.455 & 0.455 & 0.03 & 0.88 & 0.03 \end{pmatrix}.$$

Usando la nomenclatura introducida un poco antes, x es un vector propio de P de valor propio 1. Las preguntas que surgen naturalmente ya las hemos hecho: ¿hay un tal vector propio? ; ¿es único?

Las respuestas afirmativas a estas preguntas tienen un poderoso basamento teórico. El **teorema de Perron - Frobenius** sobre los valores propios de matrices con entradas reales no negativas es una pieza clave para mostrar que el método usado por PageRank funciona. En su versión original de Perron (1907) el teorema expresa que el valor propio de mayor valor absoluto de una matriz (con entradas) positiva(s) es positivo y su espacio propio es generado por un vector propio de coordenadas del mismo signo. Frobenius (1908, 1912) extendió estos resultados a matrices no negativas.

Si calculamos directamente⁴ el vector propio, se obtiene el resultado (aproximado) indicado en la Sección 3
 $x = (0.1003570039, 0.1655458921, 0.2081976187, 0.2069679755, 0.3189315099).$

El Teorema de Perron-Frobenius incluye un resultado aún más vigoroso. Si se tiene cualquier *vector de probabilidad*

$$p = (p_1, p_2, \dots, p_n) \text{ donde } 0 \leq p_i \leq 1, \quad \sum_{i=1}^n p_i = 1$$

y aplicamos sucesivamente la matriz P :

⁴ Utilizando la función vector propio de algún programa de cálculo, Matlab por ejemplo.

$$p^1 = Pp, \quad p^2 = Pp^1, \dots, \quad p^k = Pp^{k-1},$$

entonces el vector p^k se irá aproximando al vector propio x . Este resultado es central a la hora de implementar computacionalmente el cálculo. Google utiliza $d=0.85$ y realiza entre 50 y 100 iteraciones. En nuestro caso, si tomamos $p=(1,0,0,0,0)$ resultará $p^7 = (0.03, 0.03, 0.455, 0.03, 0.455)$

Si tomamos $k = 10$ y $k=11$, tenemos:

$$p^{10}=(0.0993435488, 0.1670064946, 0.2099465558, 0.2052188339, 0.3184845673)$$

$$p^{11}=(0.1009777602, 0.1653559411, 0.2075769493, 0.2084545724, 0.3176347772)$$

que está muy cerca del vector propio (aproximado) calculado anteriormente e indicado en la Sección 3.

Si el número de páginas es N , consideramos la matriz P construida de manera análoga a lo hecho en nuestro ejemplo. La matriz es enorme. Esta matriz tiene todas sus entradas no negativas y la suma de los elementos de cada columna da uno; se dice que es una *matriz de Markov*. De acuerdo con lo expresado anteriormente se trata de encontrar un vector x tal que $x=Px$.

Se prueba que si la matriz es de Markov y $\sum_i x_i = 1$, entonces x es único. El elemento x_j de x es el PageRank de la página j .

7. REFERENCIAS

- [BDJ] M. Berry, Z. Drmac & E. Jessup, *Matrices, Vector Spaces and Information Retrieval*, SIAM Review 41 (1999), 335-362.
- [BP] Sergey Brin & Lawrence Page, *The anatomy of a large scale hypertextual web search engine*, Computer Networks and ISDN Systems, 33 (1998), 107-117.
- [Bra] Daniel Brandt, *PageRank: Google's original sin*, <http://www.google-watch.org/pagerank.html>
- [Ec] *How good is google?* The economist, print edition, October 30th, 2003,
- [EF] *A Survey of Google's PageRank*, <http://pr.efactory.de/>
- [Gr] Juan-Miguel Gracia, *Álgebra Lineal tras los buscadores de Internet*, <http://www.vc.ehu.es/campus/centros/farmacia/deptos-f/depme/gracia1.htm>
- [Ka] Jerry Kazdan, *Solving Equations, An elegant Legacy*, American Math. Monthly, 105 (1998), 1-21. Versión expandida en <http://www.math.upenn.edu/~kazdan>
- [La] Pierre Lazuly, *El mundo según Google*, Le Monde diplomatique-el Dipló, edición cono sur, Octubre 2003, 36-37.
- [MC] C.R. MacCluer, *The many proofs and applications of Perron's Theorem*, SIAM Review 42 (2000), 487-498.
- [MM] R. Markarian & N. Möller, *La importancia de cada nodo en una estructura de enlaces*, Boletín de la Asociación Matemática Venezolana, 11(2004) se publicará en diciembre 2004. Disponible on-line. En <http://www.fing.edu.uy/~nmoller>
- [Mo] Cleve Moler, *The World's Largest Matrix Computation. Matlab News and notes, Cleve's corner*, <http://www.mathworks.com/company/newsletter/clevescorner/>
- [Se] E. Seneta, *Non-negative Matrices and Markov Chains*, 2nd. Edition. Springer, 1981.
- [Wi] Herbert Wilf, *Searching the web with eigenvectors*, <http://www.math.upenn.edu/~wilf/>
- [WSJ] Wall Street Journal, 26 de Febrero de 2003.

t).