Como cuantificar la importancia individual en una estructura de enlaces: Google-PageRank

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

Resumen

Al buscar información en Internet, cobra importancia, debido a su tamaño, cómo se presentan o se ordenan la variedad de 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" y permite ordenar los resultados de la búsqueda. Este método fue lo que popularizo al buscador Google. El modo en que se enlazan la páginas de la Red, determina una matriz cuyos vectores propios tienen propiedades que permiten utilizar sus componentes como una medida de "relevancia". Nuestro objetivo es ilustrar cómo son utilizadas ciertas nociones básicas del Algebra Lineal en estos cálculos.

Introducción

La siguiente cita, extraída del periódico Le Monde [La], ilustra en términos generales las ideas que profundizaremos aplicando de algunos elementos del Algebra Lineal.

> "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."

El mismo artículo continua dando más detalles

> " Para evaluar la 'pertinencia' de las páginas existentes en internet, Brin y Page inventaron el 'Page Rank', una escala de valores propia de Google. En la misma, 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."

Al buscar material en Internet se plantean simultáneamente dos problemas:

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

En este trabajo, nos concentraremos en un aspecto matemático relacionado con el segundo punto: analizaremos uno de los elementos que utiliza el buscador Google para ordenar los datos que se relacionan con nuestra búsqueda. Aunque el primer problema tiene también gran relación con elementos del Algebra Lineal no será tratado aquí; una referencia breve es [BDJ].

Los resultados iniciales de nuestra búsqueda suelen ser muchas páginas de direcciones relacionadas con el tema, pero pocas veces miramos más allá de las primeras. Por ello es muy útil un procedimiento que ordene los resultados de acuerdo a la "relevancia"que tienen las páginas.

Es allí donde interviene uno de los principales elementos introducidos por el Google en 1998, el PageRank [BP]: "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."

Expresándolo de manera un tanto simplificado, lo que buscamos es que *la importancia de cada página sea proporcional a la suma de las importancias de todos los sitios que enlazan con ella.* Matemáticamente, si llamamos K a la constante de proporcionalidad, y en cada igualdad el lado derecho es la suma de la importancia de todos los sitios que enlazan a la página del lado izquierda, tendremos un sistema de ecuaciones del tipo

$$\begin{array}{l} x_1 = K(x_{14}+x_{97}+x_{1002}) \\ x_2 = K(x_{11104}+x_{20006}) \\ \cdots = \cdots \end{array}$$

Sea A la matriz cuya entrada a_{ij} es 1 si el sitio j tiene un enlace con la página i y 0 caso contrario. Esta matriz nos permite reescribir las ecuaciones anteriores en la forma

$$x_i = K \sum_{j=1}^{n} a_{ij} x_j.$$

Entonces, el problema de hallar vectores $x = (x_1, x_2, \cdots, x_n)$ que satisfacen esa igualdad se transforma en halla x tal que $Ax = \frac{1}{K}x$, que es un problema de valores y vectores propios de una matriz que toma en cuenta la estructura de vínculos (*links* en inglés) de la Red.

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. Este resultado es central a la hora de implementar computacionalmente el cálculo.

En los años 50, ya se había observado el papel de este vector propio asociado al mayor valor propio de una matriz positiva, para obtener un ordenamiento [Ke]. Su vigencia actual se debe a su aplicación a Internet y a la posibilidad de implementar el cálculo para matrices muy grandes

Este trabajo, por su destino original (artículo dirigido a alumnos de un segundo curso de Algebra Lineal) contiene detalles conocidos por toda aquella persona con una formación profesional en el cual el teorema de Perron-Frobenius sea utilizado frecuentemente. Por lo tanto puede ser leído de varias maneras; quienes estén interesados en la descripción general del PageRank y sus implicaciones pueden leer las Secciones 1, 2, 4 y 5. Quienes estén interesados en esos asuntos y en el planteamiento y solución de los principales problemas matemáticos deben leer además la Sección 3 y principalmente el Apéndice.

1. Algo de Historia

La Red ha crecido en una forma vertiginosa. Hagamos un poco de historia para situar el contexto de invención del procedimiento (algoritmo) utilizado por Google.

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

En principio [Pe], todo motor de búsqueda está compuesto de 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 es através de ella que consultan la base de datos. El robot de indexación es el encargado de "navegar"la Web colectando toda la información que este pueda procesar y almacenándola en la base de datos para su posterior consulta.

Muchos motores desarrollaban tecnologías que permitían restringir la búsque-da. Estas restricciones empleaban argumentos lógicos que no eran de manejo sencillo. Yahoo 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, y los que funcionaban bien eran un gran negocio: Yahoo se vendió en una abultada cifra. Los algoritmos de búsqueda recibían un gran impulso y a pesar de ello no se simplificaba el hallar lo deseado.

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 posgrado donde se definía la "importancia" de una página web tomando en cuenta los enlaces que recibe. 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 "relevancia"). Algunas de éstas serán comentadas más adelante.

2. Como ordenar las páginas de la Red.

Estando en una página web A determinada tenemos dos números importantes:
- cantidad de vínculos entrantes = cantidad de páginas que tienen un vínculo hacia la página A;
- cantidad de vínculos salientes.

Las páginas web varían mucho en el número de vínculos entrantes que poseen. Generalmente las páginas que tienen muchos vínculos entrantes son más importantes que las que sólo tienen unos pocos.
Sin embargo, hay muchos casos en los cuales 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 A utilizando la Red. Se define el PageRank por:

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

donde:

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

Como la suma de esos números sobre todas las páginas web, da uno 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:

- está definido para cada página y es determinado por los PageRanks de las páginas que tienen un vínculo dirigido hacia ella;
- los sitios que enlazan a la página A no influyen uniformemente; 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.

En sus explicaciones Brin y Page 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 tener en cuenta el contenido de los mismos (esto ha suscitado comentarios y modelos alternativos ver [DR]). Cuantificando esos comportamientos posibles, se supone que seguirá un enlace de la página en que está con probabilidad d, o que salta a cualquier página con probabilidad $1-d$.

La definición del PageRank establece **un** procedimiento para determinar una probabilidad de que un usuario aleatorio llegue a la página web A. El navegador aleatorio visita una página web con una cierta probabilidad

proporcional al PageRank de la página. La probabilidad de que el usuario elija un vínculo depende de los vínculos que puede elegir en la página en que está.
El seguimiento de los vínculos está indexado probabilísticamente por el factor de amortiguamiento d. Parece razonable suponer que $d > 1/2$, o sea, estando en una página, se tiende a usar más los vínculos que allí están, que hacer una nueva elección al azar. En la Sección 3 profundizaremos en el significado y el uso de d.

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 $\frac{1-d}{N}$.

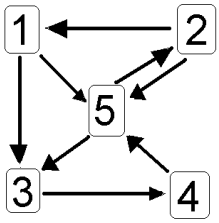
Vamos a ver que, por la naturaleza de la definición, es posible utilizar un algoritmo iterativo que aproxima los valores de PageRank. O sea, a cada página se le asigna un valor inicial y se realizan iteraciones que modifican sucesivamente estos valores iniciales. Esto es, a partir de distribuciones iniciales 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 propio que mediante otros procedimientos.

Algunas preguntas surgen naturalmente; ¿Por qué este procedimiento funciona? ¿Será que este procedimiento lleva a dar a cada página un valor único, su PageRank? Explicaremos en detalle como se realiza este cálculo en el ejemplo de la siguiente Sección.

Las respuestas afirmativas, en general, incluyen el uso de una versión del teorema de Perron-Frobenius que se dará en el Apéndice.

3. Un ejemplo

Veamos ahora como es el procedimiento recursivo. Comenzamos con una versión simplificada del problema dada por el siguiente diagrama.



Supongamos que 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 de una manera más compacta, intentando utilizar la nomenclatura probabílistica relacionada con la distribución estacionaria de una cadena de Markov [1]. Llamamos $\pi_i = PR(i)$ al PageRank de la página i:

$$\pi_1 = \tfrac{1-d}{5} + d\left(\tfrac{\pi_2}{2}\right),$$

$$\pi_2 = \tfrac{1-d}{5} + d\left(\tfrac{\pi_5}{2}\right),$$

$$\pi_3 = \tfrac{1-d}{5} + d\left(\tfrac{\pi_1}{2} + \tfrac{\pi_5}{2}\right),$$

$$\pi_4 = \tfrac{1-d}{5} + d\left(\pi_3\right),$$

$$\pi_5 = \tfrac{1-d}{5} + d\left(\tfrac{\pi_2}{2} + \pi_4\right).$$

Si definimos la matriz:

$$P = \tfrac{1-d}{5}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$
$$+d\begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1 & 0 \end{pmatrix}$$

[1] Ver el capítulo 5 de [Ha].

y $\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{pmatrix}$ y utilizando que $\sum_{i=1}^{5} \pi_i = 1$ podemos resumir las 5 ecuaciones en[2]

$$\boxed{\pi = P\pi.}$$

Modelo de navegación[3]
Obtendremos la matriz P recurriendo a las explicaciones dadas en la sección anterior para justificar la definición. Resumamos la estructura de vínculos en lo que llamamos la **matriz de conectividad** A definida por

$$a_{ij} = \begin{cases} 1 \text{ si hay un vínculo de la página j a la i,} \\ 0 \text{ si no hay un vínculo de la página j a la i.} \end{cases}$$

En el caso del ejemplo

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

Supongamos que el usuario tiene los siguientes dos modos de navegación:

1. Elige una página al azar.

2. Sigue los vínculo de la página en la que está.

Elegimos un número d, $0 < d < 1$; la probabilidad del modo 2. Queremos saber lo siguiente: estando en una página determinada ¿cuál es la probabilidad de que en el próximo paso esté en otra determinada página?

Para esto introducimos otra matriz P, llamada **matriz de transición** cuya entrada p_{ij} es la probabilidad de que estando en la página j pase al sitio i. Tenemos que:

$$p_{ij} = \begin{cases} \frac{d}{\text{vínculos salen de j}} + \frac{1-d}{\text{total páginas}} & \text{si hay un vínculo de j a i} \\ \frac{1-d}{\text{total páginas}} & \text{si no hay vínculo de j a i} \end{cases}$$

Observe que la matriz P se puede obtener a partir de la matriz de conectividad A de la siguiente manera:

- $C(T_j)$= vínculos que salen de la página j $= \sum_{i=1}^{5} a_{ij}$.
- Dividimos la columna j de A por $C(T_j)$.
- Formamos una nueva matriz C con la columnas del paso anterior.
- Entonces: $P = \frac{(1-d)}{5} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix} + dC$.

Proceso de Iteración
Llamemos vector de probabilidades a

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_5 \end{pmatrix}$$

donde $0 \leq p_j \leq 1$ y $\sum_{j=1}^{5} p_j = 1$. El número p_j es la probabilidad de estar en la página j.

Si $p^{(k)}$ es el vector de probabilidades en el k-ésimo paso de la navegación, tendremos que

$$p^{(k+1)} = Pp^{(k)}.$$

Por ejemplo si comenzamos en la página 1:

$$p^{(0)} = (1, 0, 0, 0, 0)$$

[2]En lenguaje probabilístico suele ser más común llamar P a la transpuesta de nuestra matriz y llegar a $\pi = \pi P$.
[3]Si ha estudiado cadenas de Markov es posible que las siguientes explicaciones le resulten elementales y rudimentarias.

después del primer paso $p^{(1)} = Pp^{(0)}$ es el vector

$$\begin{pmatrix} \frac{(1-d)}{5} \\ \frac{(1-d)}{5} \\ d/2+\frac{(1-d)}{5} \\ \frac{(1-d)}{5} \\ d/2+\frac{(1-d)}{5} \end{pmatrix}.$$

De la misma manera, al segundo paso $p^{(2)} = Pp^{(1)} = P^2p^{(0)}$ y podemos continuar para obtener que $p^{(k)} = Pp^{(k-1)} = P^kp^{(0)}$. Observe que todos los vectores $p^{(k)}$ son de probabilidad.

Lo que nos interesa son las probabilidades a largo plazo; o sea, nos preguntamos si los vectores de probabilidad $p^{(k)} = P^kp^{(0)}$, $k = 1, 2, \ldots$ convergen a algún vector de equilibrio π **independientemente** del vector de probabilidades inicial $p^{(0)}$. Si eso sucede, entonces en particular

$$\pi = \lim_{k\to\infty} P^{k+1}\pi = \lim_{k\to\infty} PP^k\pi = P(\lim_{k\to\infty} P^k\pi) = P\pi.$$

Por lo tanto π es un vector propio asociado al valor propio 1. Además por la independencia del vector inicial, si consideramos los vectores e_j de la base canónica tenemos que las columnas de P^k son

$$P^ke_j \to \pi,$$

lo que implica que $P^k \to P_\infty$ donde cada columna de P_∞ es el vector π.

Convergencia en norma uno[4]**.** Es importante tener en cuenta que los límites que aparecen en este trabajo se refieren a que la norma uno del vector diferencia tiende a cero. Esto significa que si $v_j^{(t)}, v_j$ son las coordenadas j-ésimas de $v^{(t)}$ y v (vectores de $\mathbb{R}^n$), respectivamente, entonces $\lim_t v^{(t)} = v$ significa que

$$||v^{(t)} - v||_1 = \sum_{j=1}^{n} |v_j^{(t)} - v_j| \to 0 \quad \text{cuando } t \to \infty.$$

Veamos cómo hacemos el cálculo **recursivamente** en nuestro ejemplo utilizando $d = 0{,}85$[5]

1. Comenzamos con un vector de probabilidad inicial

$$p^{(0)} = \begin{pmatrix} 0{,}2 \\ 0{,}2 \\ 0{,}2 \\ 0{,}2 \\ 0{,}2 \end{pmatrix},$$

2. Calculamos

$$\begin{aligned} p^{(1)} = Pp^{(0)} &= \begin{pmatrix} ,03 & ,455 & ,030 & ,030 & ,030 \\ ,03 & ,03 & ,03 & ,03 & ,455 \\ ,455 & ,03 & ,03 & ,03 & ,455 \\ ,03 & ,03 & ,88 & ,03 & ,03 \\ ,455 & ,455 & ,03 & ,88 & ,03 \end{pmatrix} \begin{pmatrix} 0{,}2 \\ 0{,}2 \\ 0{,}2 \\ 0{,}2 \\ 0{,}2 \end{pmatrix} \\ &= \begin{pmatrix} ,115 \\ ,115 \\ ,2 \\ ,2 \\ ,37 \end{pmatrix}. \end{aligned}$$

Miramos cuán próximos están:

$$\begin{aligned} \delta_1 &= ||p^{(1)} - p^{(0)}||_1 \\ &= |,115 - 0{,}2| + |,115 - 0{,}2| \\ &\quad +|,2 - 0{,}2| + |,2 - 0{,}2| + |,37 - 0{,}2| \\ &= ,34. \end{aligned}$$

3. Calculamos $p^{(k)} = Pp^{(k-1)}$, hasta que estén suficientemente próximos, o sea $\delta_k < \varepsilon$.[6]

4. Si ε es muy pequeño, la componente i del vector $p^{(k)}$ será una buena aproximación al PageRank de la página i.

En el ejemplo si tomamos k=11, llegamos a:

$$p^{(10)} = \begin{pmatrix} ,09934354886 \\ ,1670064946 \\ ,2099465558 \\ ,2052188339 \\ ,3184845673 \end{pmatrix} \quad p^{(11)} = \begin{pmatrix} ,1009777602 \\ ,1653559411 \\ ,2075769493 \\ ,2084545724 \\ ,3176347772 \end{pmatrix}$$

[4]En lenguaje probabilístico convergencia en variación total.

[5]Sólo aparecen los resultados de los cálculos; fueron calculados computacionalmente

[6]El problema de estimar k en función de ε tiene importancia a la hora de implementar el cálculo; esta estimación involucra a otro valor propio de P [HK].

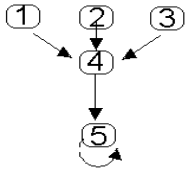
$$\Rightarrow \quad \delta_{11} = ||p^{(11)} - p^{(10)}||_1 = ,00973989994.$$

Si calculamos directamente[7] el vector propio, se obtiene un resultado muy cercano al anterior

$$\begin{pmatrix} ,1003570039 \\ ,1655458921 \\ ,2081976187 \\ ,2069679755 \\ ,3189315099 \end{pmatrix}.$$

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**.) Si vuelve al ejemplo anterior verá que aparece una división por 0 al definir la matriz P. En ese caso se calcula el de las páginas 1, 2, 3, 4 y después con esos números el de la 5. Esto es un fenómeno presente muchas veces en el cálculo del PageRank real, por ejemplo debido a enlaces a páginas que no han sido todavía descargadas por las "arañas" del Google (éstas aparentan no poseer enlaces salientes).

4. Google en serio

Se considera el conjunto W de páginas que se pueden alcanzar a partir de una página en Google. Sea N el número de páginas en W, este número varía con el tiempo, (en mayo 2002 era alrededor de 2700 millones [Mo]). Consideramos la matriz $N \times N$ de conectividad de W. La matriz es enorme pero tiene una gran cantidad de ceros (en ingles, *sparse matrix*). Consideramos la matriz P construida de manera análoga a lo hecho en nuestro ejemplo. 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 en la Sección anterior se trata de encontrar un vector π tal que $\pi = P\pi$. Se prueba que si la matriz es de Markov y $\sum_i \pi_i = 1$, entonces π es único. El elemento π_j de π es el PageRank de la página j (a menos de posibles cambios de escala).

Observemos que la forma recursiva de implementar el algoritmo al realizar el cálculo no es algo menor, estamos hablando de manejar una matriz que tiene un tamaño de varios millones. En el Apéndice se mostrará porque funciona esta implementación que asigna una calificación no nula única a cada página.

En teoría sucede que toda página posee un PageRank positivo pero en el ordenamiento real se introducen como penalización una calificación nula llamada **PR 0**. Desde que se popularizo la utilización del Google los responsables (*webmaster*) de algunas sitios han intentado aumentar la calificación de sus páginas intentando manipular sus 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].

5. Consideraciones generales

En este momento, Google no sólo es el buscador más utilizado, sino que, le vende servicios a portales importantes: Yahoo, AOL, etc. 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.

El 27 de 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.

Por ello, cualquier algoritmo como éste, que aparenta ser objetivo, continuará siendo un aspecto importante para las búsquedas en la Red.

[7] Utilizando la función vector propio del programa de cálculo.

Google es también el único motor 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 dias. Se ha advertido, que el PageRank vigente influye el recorrido mensual realizado por Google, hace que páginas con PageRank más alto sean recorridas más rápidamente y en 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.

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. Ya hemos observado que existen quienes intentan mejorar su calificación, y que, Google trata de controlar tales comportamientos. Se han realizado experiencias exitosas que muestran las posibilidades de utilizar "artificialmente"esta característica para subir el PageRank de una página. En los términos utilizados en [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."

El artículo enfatiza aun:

> "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. La base y la sobrerepresentación de que se benefician los 'adelantados' de internet, la densidad de lazos que mantienen (sobre todo a través del fenómeno esencialmente estadounidense de los *weblogs*), designan -matemáticamente- a los actuales 'gurús' de Google. Por cierto que el sistema pasó brillantemente las pruebas en cuestiones técnicas y prácticas. Pero existen terrenos en los que la pertinencia escapa a los algoritmos."

Google, empresa, está próxima a lanzar au cotización en bolsa. Su precio se estima en varios miles de millones de dólares [Ec]. Se dice que Microsoft también estaría por lanzar su propia tecnología de búsqueda.

Apéndice: Porqué funciona el algoritmo.

Importancia de las matrices no negativas.

En este Apéndice daremos una demostración algebraica de una versión probabilística del Teorema de Perron-Frobenius.

Distintas versiones de este teorema fueron probadas en contextos totalmente abstractos, pero la importancia de la teoría de matrices no negativas se ha extendido a campos muy amplios: las teorías de probabilidad y de sistemas dinámicos, el análisis numérico, la demografía, la economía matemática y la programación dinámica. Ver, por ejemplo [MC].

Esto se debe a que diversas variables que se miden en el mundo real, interactúan a través de relaciones positivas o nulas. A su vez, una cantidad de modelos que formalizan esas interacciones son procesos iterativos lineales en que se comienza con un estado v y se evoluciona por la aplicación reiterada de una matriz A, de modo que luego de n pasos se tiene el estado $v^{(n)} = A^n v$. Muchas veces es fundamental saber cuándo este proceso converge a un estado único, cualquiera sea el estado de comienzo v. La teoría de matrices positivas responde a ésta (y muchas otras) cuestión(es).

El enfoque que haremos aquí tiene por prerequisitos algún manejo algebraico y cursos elementales de álgebra lineal. Este enfoque sencillo y directo puede hacer perder parte del "sabor probabilístico"que en profundidad tienen muchos de los contenidos. Pedimos disculpas por esta opción que no es involuntaria.

Se pueden encontrar otras pruebas y desarrollo de estos temas en [MC]; [Ha] Ch. 5; [Re] Ch. 2. Un tratado muy completo sobre matrices no negativas, que comienza con los resultados que nos interesan, es [Se].

Convergencia

Sea $B \in M_{n\times n}(\mathbb{C})$ una matriz con valor propio $\lambda_1 = 1$ de multiplicidad algebraica 1, y los demás valores propios satisfaciendo $1 > |\lambda_2| \geq \ldots \geq |\lambda_n|$. Los $\lambda_i, i \geq 2$ se pueden repetir. El uno es lo que se llama un **valor propio dominante**.

Si existe una base de vectores propios v_i, $i = 1, \ldots, n$, entonces $v = \sum_{i=1}^{n} \alpha_i v_i$ y

$$B^k(v) = \alpha_1 v_1 + \sum_{i=2}^{n} \alpha_i \lambda_i^k v_i,$$

donde sabemos que $\lim_k \lambda_i^k = 0$ para $i = 2, \ldots, n$.

Si no hay una base de vectores propios, o sea si la multiplicidad algebraica de alguno de los $\lambda_i, i \geq 2$; es diferente de la multiplicidad geométrica, entonces habrán elementos w_1, w_2 de una base (de Jordan) que satisfacen $Bw_1 = \lambda_i w_1 + w_2$, $Bw_2 = \lambda_i w_2$. Un breve cálculo permite deducir que $B^k w_1 = \lambda_i^k w_1 + k\lambda_i^{k-1} w_2$. Este sencillo cálculo es generalizable a todas las situaciones que se pueden presentar al tomar una base de Jordan.

En la matriz de Jordan se tienen los llamados bloques de Jordan que son de la forma $\lambda I + N$ donde λ es un valor propio (real o complejo) I es una matriz identidad, digamos que s por s, y N una matriz "nilpotente"todas cuyas entradas son cero, excepto la linea subdiagonal $(a_{i,i-1})$ que está formada por unos que verifica $N^s = 0$. Entonces $(\lambda I + N)^k = \sum_{i=0}^{s} \lambda^{k-i} C_i^k N^i$, donde C_i^kson las combinaciones de k elementos tomados de a i.

Por tanto $B^k v$ será combinación lineal de vectores de la base, cuyos coeficientes –con excepción del correspondiente al vector propio asociado al valor propio 1– tienden a cero cuando[8] $k \to \infty$ por tener cada valor propio –distinto del primero– módulo menor que uno.

Así, sea la matriz B diagonalizable o no, si el uno tiene multiplicidad algebraica uno y las demás raíces del polinomio característico tienen módulo menor que uno podemos garantizar que

$$\lim_k B^k v = \alpha_1 v_1.$$

La igualdad anterior es la que nos permite realizar el cálculo iterativo cuando $\alpha_1 \neq 0$ porque la aplicación sucesiva de la matriz B a cualquier vector con $\alpha_1 \neq 0$ converge a múltiplos de un mismo vector (un vector propio de valor propio 1). El teorema de Perron-Frobenius nos permitirá también elegir vectores con los que comenzar el proceso, con $\alpha_1 \neq 0$, para los que la convergencia no será al vector nulo.

Matrices de Markov. Resultados principales.

Decimos que una matriz B es **positiva** si todos sus elementos son números positivos. Lo notamos $B > 0$. Si $X \in M_{n\times 1}$ $(\mathbb{R})$(o $M_{1\times n}(\mathbb{R})$) es positivo decimos que X es un **vector positivo**. Dadas dos matrices A, B del mismo tamaño decimos que $A > B$ si $A - B$ es una matriz positiva. Tenemos definiciones análogas para $A \geq 0$ (no negativo) si sustituimos positivos por elementos mayores o iguales a cero. Diremos que $p \in \mathbb{R}^n$ es un **vector de probabilidad** si es no negativo y la suma de sus componentes es uno.

Definición 1 (Matrices de Markov). *Dada una matriz positiva $M \in M_{n\times n}(\mathbb{R})$ decimos que es de* **Markov** *si la suma de los elementos de cada columna es uno ($\sum_{i=1}^{n} m_{ij} = 1$, $\forall j = 1, \ldots n$.)*

Se probará que las matrices de Markov satisfacen las propiedades de la matriz B anterior: tienen un valor propio uno con multiplicidad algebraica uno y todos los demás valores propios con módulo menor que uno. También se probará que si el vector v es de probabilidad el valor α_1 antes referido es distinto de cero. Gran parte de estas demostraciones se harán usando la matriz traspuesta M^T de la matriz de Markov, que satisface que los elementos de sus filas suman uno, y tiene los mismos valores propios que M (por tener el mismo polinomio característico).

En el transcurso de la demostración de los puntos principales se demostrarán otros resultados interesantes. Resumimos todos los resultados en el siguiente

Teorema. *Si M es una matriz de Markov, entonces todos sus valores propios tienen módulo menor o igual que uno y sólo uno de ellos, el uno, tiene módulo uno. La multiplicidad algebraica del uno es uno y es el único que un vector propio no nulo es positivo.*
Para cualquier vector de probabilidad p, $M^k p$ converge en la norma 1 [9] *al único vector de probabilidad que es vector propio del valor propio uno.*

Como ya se vió estas condiciones aseguran la convergencia de $M^k v$ (para cualquier v) a un múltiplo del vector propio v_1 del valor propio uno. Comencemos probando que si M es de Markov y p es de probabilidad también lo es $M^k p$. Alcanza con probar que Mp es de probabilidad. En efecto , la suma de las componentes de Mp es

$$\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} p_j = \sum_{j=1}^{n} p_j \sum_{i=1}^{n} m_{ij} = \sum_{j=1}^{n} p_j 1 = 1.$$

[8] Recuerde que $\lim_k \lambda_i^k k^s = 0$.
[9] Ver definición en Sección 3.

La segunda de las igualdades es consecuencia de que M es de Markov (ver definición). Entonces las iteraciones de un vector de probabilidad convergerán a un vector de probabilidad, que será el **único** vector probabilidad que es vector propio del valor propio uno. Por tanto el vector límite no es nulo y se muestra, de pasada, que al escribir un vector de probabilidad v como combinación lineal de los vectores de la base de Jordan, resultará $\alpha_1 \neq 0$.

Convergencia de M^T.

Probaremos ahora que para cualquier $w \in \mathbb{R}^n$, $(M^T)^k w$ converge cuando $k \to \infty$. Esta prueba será independiente de la estructura de valores y vectores propios de M^T; sólo utilizará el hecho de que M es de Markov. Necesitaremos el siguiente resultado que presentamos en forma de ejercicio con sugerencia.

Ejercicio: Sean $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, $0 < \gamma < 1/2$ tales que: $0 < \gamma \leq c_j$, $c_1 + \ldots + c_n = 1$, entonces el promedio ponderado de los números $w_1, \ldots w_n$ se define como $\overline{w} = c_1 w_1 + \ldots + c_n w_n$. Sean w_{min} y w_{max} los valores mínimo y máximo de los $w's$. Entonces el promedio $\overline{w}$ satisface:

$$\gamma w_{max} + (1-\gamma) w_{min} \leq \overline{w} \leq (1-\gamma) w_{max} + \gamma w_{min}, \text{ y } w_{min} \leq \overline{w} \leq w_{max}.$$

Se sugiere, para probar las primeras desigualdades, hacerlo por inducción completa en n, suponiendo, por ejemplo, que al pasar de n a $n+1$, se agrega el w_{max}.

Para aplicar este resultado a nuestro problema, consideramos $\gamma = \text{mín}\{m_{ij}\}$ (si $n > 2$, resulta $\gamma < 1/2$). Como las filas de M^T suman 1, los elementos de $z = M^T w$ son promedios ponderados de los elementos de w. El resultado del ejercicio nos da estimativas para las componentes máxima y mínima de z,

$$z_{max} \leq (1-\gamma) w_{max} + \gamma w_{min},$$
$$\gamma w_{max} + (1-\gamma) w_{min} \leq z_{min}.$$

Esas desigualdades implican que

$$w_{min} \leq z_{min} \leq z_{max} \leq w_{max} \ ; \quad z_{max} - z_{min} \leq (1-2\gamma)(w_{max} - w_{min});$$

como $0 < (1-2\gamma) < 1$ la diferencia entre el valor máximo y mínimo de la iteración resulta una contracción. Por tanto los vectores resultantes de la aplicación sucesiva de M^T convergen en la norma 1 a un vector no nulo con todas sus componentes iguales y distintas de cero si $w_{max} > 0$. Si se comienza el proceso tomando como vector w, cualquier vector con todas sus componentes iguales se observa que este vector z_1 es un vector propio asociado al valor propio uno.

Este es el único vector propio asociado al valor propio uno porque si hubiera otro $z_2 \neq z_1$, resultaría $(M^T)^m z_2 = z_2$, y convergería a z_1. Absurdo.

Valores propios

Veremos ahora que el uno tiene multiplicidad algebraica uno. Si su multiplicidad fuera $k > 1$, tal como ya fuera observado al principio de este Apéndice, resultará que la matriz de Jordan tendrá un bloque $I + N$ con $N^k = 0$. Por tanto $(I+N)^m v = \sum_{i=0}^{k} C_i^m N^i$ que no puede converger a un vector de coordenadas acotadas puesto que las combinaciones $C_i^m, 0 \leq i \leq k$ van para infinito con m.

Las mismas expresiones que figuran al principio de este Apéndice, al introducir la matriz de Jordan, muestran que no pueden haber valores propios con módulo mayor que uno, pues $\lambda^{m-i} \to \infty$ cuando $m \to \infty$.

Por último, debemos probar que no hay valores propios complejos de módulo 1. Si existiera uno$\lambda = a + bi$, $b \neq 0$ con argumento $\varphi \neq 0$, $|\lambda| = 1$, la matriz de Jordan real tendrá una submatriz de la forma $J = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$. Entonces $J^m = |\lambda|^m \begin{pmatrix} \cos m\varphi & \sin m\varphi \\ -\sin m\varphi & \cos m\varphi \end{pmatrix}$. Por tanto $M^m v$ tendrá dos componentes que no convergerán, porque las que correspondan a esa submatriz estarán dependiendo del valor de $m\varphi$ (si $\varphi \neq 0$, los senos y cosenos correspondientes convergen a por lo menos dos valores distintos al crecer m. Obsérvese que si φ es irracional, los senos y cosenos convergen a todos los valores entre -1 y 1).

Los resultados necesarios para asegurar la convergencia de $M^m v$ cuando v es un vector de probabilidad ya han sido probados. Ahora daremos otras pruebas de los mismos resultados, y completaremos el resultado faltante.

Otra demostración

El teorema de Gershgorin nos permite, sin calcular explícitamente los valores propios, tener una idea de su valor: *Los valores propios de una matriz* $\{a_{ij}\}$ *se encuentran en los círculos del plano complejo de centro* a_{ii} *y radio* $\sum_{i\neq j} a_{ij}$. Como nuestra matriz es de Markov los centros son $m_{ii} < 1$ y los radios $1 - m_{ii}$, por lo que los círculos que contienen los valores propios están todos dentro del círculo de centro en el origen y radio 1. Todos esos círculos contienen el punto (1,0) de lo que se deduce que el único valor propio de módulo 1 es el 1.

Observación: Podemos tener matrices con valor propio 1 pero que M^k no converja o que el límite $M^k p$ dependa de p. Ejemplos de esto son (en dimensión 2) $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ o la matriz identidad.

Sea M una matriz de Markov. Consideramos

$$r = \sup\{\lambda \geq 0 : \ Mx \geq \lambda x \text{ para algún } 0 \neq x \geq 0\}.$$

Como el uno es valor propio de M existe un vector propio $z = (z_1, \ldots, z_n)$. Si llamamos $M^{(j)}$ a la columna j de Mla igualdad $Mz = z$ la podemos escribir

$$\sum_{j=1}^{n} z_j M^{(j)} = z.$$

Si tomamos valor absoluto a ambos lados de la igualdad, utilizamos la desigualdad triangular y llamamos $|z|$ al vector $(|z_1|, \ldots, |z_n|)$ obtenemos:

$$M|z| \geq |z|.$$

De donde concluimos que $r \geq 1$.

Mostraremos que $r = 1$. Para ello alcanza con probar que r es valor propio de M pues los valores propios de M tienen módulos menores o iguales a uno. a

Afirmación: r es valor propio de M.
Sea $0 \neq \xi \geq 0$ tal que $M\xi \geq r\xi$. Si $M\xi \neq r\xi$ tenemos que el vector $0 \neq y = M\xi - r\xi \geq 0$ como M es positiva se cumple que $My > 0$ y por lo tanto existe $\varepsilon > 0$ tal que $My \geq \varepsilon M\xi$ o sea

$$M(M\xi - r\xi) \geq \varepsilon M\xi \Rightarrow M(M\xi) \geq (r + \varepsilon)M\xi.$$

Considerando el vector $M\xi$ la desigualdad anterior contradice la definición del r. Entonces $M\xi = r\xi$ como queríamos mostrar. Obsérvese que como $M > 0$ llegamos a que el vector $\xi = (\xi_1, \ldots, \xi_n)$ **es positivo**.

Afirmación: El uno es el único valor propio asociado con un vector propio z con todas sus componentes ≥ 0.
Este enunciado sólo tiene sentido para valores propios reales porque los valores propios no reales (complejos) deben tener vectores propios con algunas o todas sus coordenadas complejas. Sea $z = (z_1, \ldots, z_n)$ un tal vector propio de M, con valor propio λ. Algún z_i deber ser mayor que cero (¿Porqué?). Sea

$$\alpha = \text{mín}\left\{\frac{z_i}{\xi_i}, \ \text{ con } \ z_i \neq 0\right\}.$$

De la definición de α vemos que existe algún p, $1 \leq p \leq n$ con $z_p = \alpha\xi_p > 0$. Como $\lambda z = Mz \geq \alpha M\xi = \alpha\xi$ si miramos la componente p-ésima $\lambda z_p \geq \alpha\xi_p = z_p \Rightarrow \lambda \geq 1$.

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

[DR] Pedro Domingos & Matthew Richardson, *The intelligent surfer: probabilistic combination of link and content information in PageRank*. Advances in Neural Information Processing Systems **14** (2002).

[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, *Algebra Lineal tras los buscadores de Internet.*
http://www.vc.ehu.es/campus/centros/farmacia/deptos-f/depme/gracia1.htm

[Ha] O. Haggstrom, *Finite Markov Chains and Algotihmic Applications*, Cambridge University Press, 2002.

[HK] T. Haveliwala and S. Kamvar, *The Second Eigenvalue of the Google Matrix.* A Stanford University Technical Report http://dbpubs.stanford.edu:8090/pub/2003-20

[Ka] Jerry Kazdan, *Solving Equations, An elegant Legacy.* American Math. Monthly, **105** (1998), 1-21. Versión expandida en http://www.math.upenn.edu/~kazdan

[Ke] M. Kendall, *Further contributions to the theory of paiRed comparisons.* Biometrics **11** (1955), 43-62.

[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.

[Mo] Cleve Moler, *The Worlds Largest Matrix Computation. Matlab News and notes, Cleves corner.*
http://www.mathworks.com/company/newsletter/clevescorner/

[Pe] *Motores de consulta.*
http://librosdigitales.net/eureka/eureka0903/motores.htm

[Re] S. Resnick, *Adventures in Stochastic Processes*, Birkhauser, 1992.

[Se] E. Seneta, *Non-negative Matrices and Markov Chains.* 2md. Edition. Springer, 1981.

[We] T. Wei, *The algebraic foundations of ranking theory.* Cambridge University, England (1952). Tésis no publicada

[Wi] Herbert Wilf, *Searching the web with eigenvectors.* http://www.math.upenn.edu/~wilf/