Google Maths

por

Beatriz Rubio Serrano

(Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza)

En los últimos años el crecimiento de usuarios de Internet ha sido exponencial, hasta el punto de que la mayoría de nosotros nos conectamos diariamente. No sólo el correo electrónico es algo indispensable en nuestra comunicación, sino que la información que nos encontramos en la red es para todos nosotros esencial. Precisamos de la red para indagar desde cosas triviales como direcciones y teléfonos, mapas, ofertas... hasta compras, búsqueda de trabajo o incluso para ampliar nuestro conocimiento. ¿Quién no ha utilizado alguna vez un buscador? O mejor dicho, ¿quién no ha utilizado el buscador Google?

Google es actualmente el buscador más popular, con infinidad de páginas censadas y da servicio a más de 250 millones de consultas diarias de forma inmediata gracias a sus miles de computadoras. Google es como un bibliotecario trabajando a la velocidad de la luz: refina la búsqueda y acaba encontrando lo que queremos.

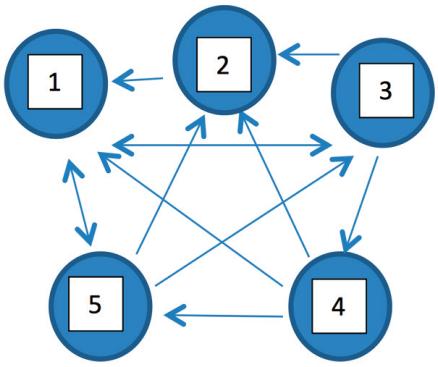
Lo que ha permitido todo este eficiente servicio de Google para indexar páginas web ha sido la informática, por supuesto, y cómo no: las matemáticas.

¿Os habéis preguntado por qué en nuestra búsqueda aparecen unas páginas antes que otras? Aquí es donde entra en juego el algoritmo de Google llamado Page Rank, inventado por los fundadores de Google, el informático Larry Page —de ahí el nombre del algoritmo— y el matemático Sergey Brin.

Veamos un ejemplo para qué podamos visualizar cómo Google a través de su algoritmo Page Rank posiciona unas páginas antes que otras.

Antes de nada debemos saber que todo Internet es a efectos de conectividad un inmenso grafo dirigido. En este grafo los nodos son las páginas web, las líneas entre los nodos son las conexiones/referencias entre páginas, y el grafo es dirigido porque dichas conexiones pueden ser en una sola dirección o en doble direccionalidad.

Bien, imaginemos que nuestra red tiene 5 páginas web, y que el grafo que representa las conexiones entre sus páginas es el siguiente:



¿Cómo interpretamos este grafo? La página web 1 enlaza a la web 5 y es enlazada por la 2, 3, 4 y 5. Así sucesivamente con el resto de páginas.

¿Y qué página es más importante? El Postulado PageRank dice lo siguiente:

La importancia x_j de la página P_j es proporcional a la suma de las importancias de las páginas que enlazan con P_j.

Es decir, será más importante mi página si además de enlazar a otras páginas es enlazada por páginas importantes. Será mucho mejor si conseguimos que nuestra página sea enlazada por Amazon.com que por una página sin apenas visitas.

El dibujo de un grafo es ilustrativo pero nosotros necesitamos operatividad, por lo que en lugar de utilizar el grafo dirigido vamos a considerar su matriz de incidencia.

La matriz de incidencia de nuestro grafo es:

$$M_I = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

En el primer modelo de Google la matriz de incidencia tenía un 1 en el lugar (i, j) si había un enlace desde p_i hasta p_j y un 0 si no lo había, pero más tarde se dieron cuenta de que si había un página con un solo enlace, este enlace valía lo mismo que cualquier otro enlace de otra página que produjese un millón de enlaces. Para solucionar esto se modificó el algoritmo Page Rank de la siguiente manera: Si hay un enlace desde desde p_i hacia p_j, en el que el lugar (i, j) de la matriz de incidencia se coloca el número . De esta forma en cada fila hay una distribución de números no negativos que suman 1, como si fuera una distribución de probabilidades sobre los nodos de internet. Este tipo de matrices se conoce como estocásticas por filas.

Por lo tanto nuestra nueva matriz de incidencia-estocástica es de la forma:

$$M_{I,E} = \begin{pmatrix} 0 & 0 & 1/2 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \end{pmatrix}$$

De momento todo pinta muy bonito, pero no hemos tenido en cuenta en nuestro grafo que una página no es enlazada por ninguna otra, y esto en la red pasa continuamente: el grafo de Internet no es fuertemente conexo, ni siquiera es conexo. Por lo tanto, lo normal es que Google se encuentre con matrices de incidencia con ceros en más de una fila, matrices no regulares, ni diagonalizables algo que trae problemas a la hora de obtener resultados. ¿Qué hace el algoritmo Pange Rank ante esta situación? Perturba la matriz estocástica de la siguiente forma:

Sea $0 < \varepsilon < 1$

$$M^{\varepsilon}_{I,E} := (1-\varepsilon) \cdot M^{\varepsilon}_{I,E} + \varepsilon/n \cdot \begin{pmatrix} 1 & \dots & 1 \\ \dots & \dots & \dots \\ 1 & \dots & 1 \end{pmatrix}$$

Donde n es el número de nodos (páginas web)

Al aplicarle ε volvemos a tener una matriz estocástica por filas, y además ahora nuestra nueva matriz $M^{\varepsilon}_{I,E}$ tiene todos sus coeficientes positivos, lo que facilita su operatividad.

Sigamos con nuestro ejemplo y calculemos la matriz perturbada de nuestro grafo.

El algoritmo Page Rank va variando varía el valor de ε, supongamos que ahora es $\varepsilon = 0{,}15$

$$M^{\varepsilon}_{I,E} = \begin{pmatrix} 0{,}03 & 0{,}03 & 0{,}455 & 0{,}03 & 0{,}455 \\ 0{,}88 & 0{,}03 & 0{,}03 & 0{,}03 & 0{,}03 \\ 0{,}313 & 0{,}313 & 0{,}03 & 0{,}313 & 0{,}03 \\ 0{,}313 & 0{,}313 & 0{,}03 & 0{,}03 & 0{,}313 \\ 0{,}313 & 0{,}313 & 0{,}313 & 0{,}03 & 0{,}03 \end{pmatrix}$$

Muy bien ya disponemos de una matriz en condiciones. Y, ¿ ahora qué?. ¿Cómo sabemos que página saldrá en la primera posición de nuestra búsqueda?

Aquí es donde entra en juego el Algebra Lineal y sus valores y vectores propios.

El algoritmo Page Rank tiene en cuenta el siguiente *Teorema:*

Si M_I es la matriz de incidencia del grafo de internet, y $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, $x_i \geq 0$, el vector de importancias, entonces se cumple $M_I^t x^t = \lambda x^t$ donde $\lambda \in \mathbb{R}$, $\lambda > 0$, es la constante de proporcionalidad.

Recordemos la definición de valor y vector de propio y así encontraremos la relación con el algoritmo Page Rank

Definición. Sea M una matriz cuadrada de orden n, $M=(a_{ij})$, tal que sus elementos a_{ij} son reales. Se llama *vector propio* de M a todo vector $X=(x_1, x_2, \ldots, x_n) \neq 0$, $x_i \in \mathbb{R}$, $i=1,2,\ldots,n$, tal que $MX^t=\lambda X^t$ con $\lambda \in \mathbb{R}$. Al número λ se le denomina *valor propio* de M asociado al vector propio X

Por lo tanto, el vector de importancias de las páginas web es un vector propio (positivo) de la matriz M_I^t, y la constante de proporcionalidad λ es el valor propio asociado a este vector.

Volviendo a nuestro ejemplo, hallemos los valores y vectores propios de nuestra matriz de incidencia-estocástica *perturbada*. Para hacer los cálculos utilizamos Wolphram Alpha y la matriz traspuesta de $M^{\varepsilon}_{I,E}$.

$$M^{\varepsilon}_{I,E} = \begin{pmatrix} 0{,}03 & 0{,}03 & 0{,}455 & 0{,}03 & 0{,}455 \\ 0{,}88 & 0{,}03 & 0{,}03 & 0{,}03 & 0{,}03 \\ 0{,}313 & 0{,}313 & 0{,}03 & 0{,}313 & 0{,}03 \\ 0{,}313 & 0{,}313 & 0{,}03 & 0{,}03 & 0{,}313 \\ 0{,}313 & 0{,}313 & 0{,}313 & 0{,}03 & 0{,}03 \end{pmatrix}$$

Ver todos los valores y vectores propios de $M^{t\varepsilon}_{I,E}$ en Wolframalpha. Notar que es una matriz diagonalizable: todos sus valores propios son simples.

De todos los posibles valores propios (aproximados) escogemos $\lambda=1$ (el único real y positivo)

Su vector propio asociado es en valor absoluto:

$$(0.67259,\ 0.363478,\ 0.463318,\ 0.194141,\ 0.403921)$$

Por lo tanto el orden en el aparecerían las páginas de nuestra red en nuestra búsqueda serian:

Página 1
Página 3
Página 5
Página 2
Página 4

Os dejamos como ejercicio que penséis, junto al grafo, la justificación del orden de importancia de las páginas.

Como ya habréis pensado, el grafo de Google es gigantesco y su matriz tiene más de un billón de entradas. Resolver una matriz de este tamaño, calcular su polinomio característico de grado billón, sus valores propios, vectores propios es muy costoso incluso para los ordenadores más potentes. Google soluciona este *problema* aplicando el método de las potencias. Aproximadamente dice lo siguiente:

Si una matriz cuadrada M es *diagonalizable* y tiene todos sus vectores $\{v_1,\ldots\ldots,v_2\}$ numeradas de tal manera que los vectores propios correspondientes cumplan lo siguiente

$\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_3|$ partiendo de $v_0 \geq 0$ tal que
$v_0 = \alpha_1 v_1 + \ldots + \alpha_n v_{n,}$ con $\alpha_1 \neq 0$, entonces se tendrá:
$M^K v_0 = \alpha_1 \lambda^K{}_1 v_1 + \ldots. + \alpha_n \lambda^K_n v_{n,}$

Luego

$$\mathrm{Lim}_{k\to\infty} \frac{M^k v}{\lambda^k} = \alpha_1 v_1$$

Es múltiplo no trivial del vector propio buscado.

Sin entrar en más detalles técnicos este es, a grandes rasgos, el método que utiliza Google para ordenar sus páginas. Hemos podido comprobar cómo la Teoría de Grafos, el Algebra Lineal y otras ramas de las Matemáticas son fundamentales en su funcionamiento. En definitiva podemos resumir: El algoritmo Page Rank de Google está basado en matemáticas. ¿Se lo contamos a nuestros alumnos?