# ¿Cómo funciona Google?

La idea es explicaros por encima cómo funciona el mecanismo de Google. Vosotros sois de la generación de Google y de Internet. Tenéis un dedo acostumbrado a hacer clic. Para mi generación, en cambio, ir a buscar información significaba que cogías la mochila e ibas a las bibliotecas a hacer fotocopias.

Esto se ha terminado. ¿Qué parte importante de Google está incorporada a nuestra vida? ¿Quién de vosotros no se ha conectado a Google en los últimos tres días para buscar cualquier tipo de información? Antes buscábamos una enciclopedia, ¿qué hacemos ahora habitualmente?

Cuando buscas una cosa te conectas a Google. Habitualmente haces Google y miras en Wikipedia. ¿No os ha pasado alguna vez? ¿Nunca os habéis preguntado por qué la Wikipedia Siempre es uno de los primeros resultados?

O por qué cuando buscáis para comprar zapatos aparecen unas marcas determinadas como Nike o lo que sea y no sale la tienda del barrio? ¿Por qué ocurre esto? Esto ocurre por el sistema que utiliza Google para hacer este tipo de rankings, lo que llamamos ordenar las páginas web cuando haces una búsqueda. Pensad que Google es muy joven, Google está en primero de bachillerato, nació en el 98, así que es como vosotros. Un poco mayor o un poco menor, pero está en primero de bachillerato.

Empezaremos hablando un podo de informática. Yo no soy informático, soy matemático, por lo tanto, os explicaré un poco como funciona Google, como busca.Y si tenéis un informático en el instituto preguntadle, porque seguro que sabe mucho más que yo.Después iremos a la parte de los algoritmos, cómo se inventan, pero claro, antes os explicaré cómo se descubrió.

La idea es cómo empieza, cuáles son los problemas, cómo se aborda, cuáles son los problemas, y al final, cuál es el resultado final.

Empezamos con la historia de Internet. Fijaos. Primeros pasos: se crea lo que sería el embrión de Internet, en el año 58, con empresas como Dell i con el MIT, el Massachusetts Institute of Technology. Empezaron a hacer cosas parecidas a redes, pero donde se desarrolló realmente fue en el CERN.

Fijaos que ya estamos hablando del año 90. La primera Internet de verdad, la World Wide Web, nació en 1990 en el CERN, en Suiza.

Fundación de Google, en el año 98: dos estudiantes d'Stanford, Sergey Brin, que es matemático, y Lawrence Page, un estudiante de computación, de informática, están preparando el trabajo de final de grado y su idea es crear una red, un sistema que permita ordenar las cosas que buscas. Y empiezan a hablar de Google. De hecho no empezó como Google, se llamaba BackRub, el nombre Google vino después.

La mitología dice que Google es una versión mal pronunciada de Gúgol, que es $10^{100}$. A partir de aquí, Google fue el nombre oficial. Pero fijaos, en el 98. Y estamos en 2015. Por lo tanto, hace muy poco y mirad la cantidad de personas que dependen de Google desde este momento. Ahora os daré una serie de datos.

Google responde cada día aproximadamente a 1 billón de consultas, a 181 países y en 145 idiomas diferentes.

Por lo tanto, imaginad la magnitud y el poder que tiene Google. Google empezó muy pequeño y ahora reúne un poder muy grande: facebook, Gmail, etc.

¿Cómo funciona Google? ¿Qué hace Google y cualquier buscador, un *search machine*, un buscador de Internet, que hay muchos: Google, Altavista, MSN, Yahoo!...? Hay muchos tipos de buscadores, muchísimos.

La manera clásica consta de 4 pasos, también llamados módulos.

Uno es el *crawler*, que se puede traducir como gatear. ¿Y qué es eso? Los famosos robots. Cuando escribís en algún lugar... a veces os dicen: "escribe estos códigos para verificar que eres realmente una persona". ¿Por qué lo hacen? Para evitar a los robots. ¿Qué son los robots? Se llaman *spiders*, es decir, arañas, y rastrean todos los enlaces que tienen las páginas web.Todos, absolutamente todos. Y van guardando toda la información. Esta información se lleva a unos servidores que están en todo el mundo. De hecho se dice que Google tiene aproximadamente, un millón de servidores en todo el mundo.

Allí van acumulando toda la información. Estos son rastreadores puros. ¿Cuál es la segunda parte? El módulo de indizar, que es como si tuvieseis a gente allí. Estos servidores van releyendo toda la información y la van clasificando, se quedan con la parte más importante. Hacen las *keywords* (palabras clave), fotografías, imágenes, textos importantes... y lo van guardando. Esta es la parte que será realmente importante para hacer después la clasificación. Lo guardan comprimido.

Fijaos que hasta aquí, el individuo, el usuario, no actúa. Aquí no entráis vosotros. ¿Dónde entráis vosotros? Aquí. En el cuarto módulo. Vosotros llegáis y hacéis la pregunta: zapatos de no sé qué. Entonces, ¿qué se hace automáticamente? ¿Qué hacen Google y los otros buscadores? Eso lo traducen a números, a números y códigos y van a buscar toda la información que tienen guardada, como si fueran a la biblioteca, de todo aquello que está relacionado con lo que habéis preguntado. Después viene el módulo más importante, el *ranking* que es ordenar esa lista.

Bien pues nosotros lo que aremos es eso, explicar cómo se hace ese ranking. Pensad que es muy importante, si tenéis una empresa, ¿qué es lo que desearíais? Que cuando te busquen en Google, aparezcas el primero, el segundo o entre los cinco primeros de la lista. Porque con esto ganaríais muchísimo, la diferencia entre el éxito y el fracaso de una empresa en Internet puede ser el lugar en el que aparezcas en la búsqueda de Google.

Muy bien, empecemos. ¿Cómo ordena Google todas estas páginas por orden de importancia? ¿Y qué haremos? Pues como se hace muchas veces en matemáticas: empezaré con un problema muy pequeño y lo escalaré para hacerlo más grande.

Imaginad que sois... cojamos cinco o seis y cada grupo tiene su página web. Después imaginad que entre estos cinco o seis hacéis enlaces entre las páginas web, por ejemplo: ve a la página de Manolito o de Pepito porque es una página muy interesante.

La idea es: cuando tengo esto de aquí y lo miro desde fuera, ¿cómo os ordenaré por importancia? Empecemos. La primera opción es por popularidad. ¿Quién es el más popular de los cinco? Seguramente será quien salga más veces enlazado y cuando haga la ordenación será el primero de la lista.

Muy bien, veamos si funciona o no funciona. Cojamos una red muy pequeña. Pensad que yo cogeré una red muy pequeña, una red de cinco o seis puntitos, pero la cantidad de páginas web que Google indexa actualmente es del orden de 2,7 billones con b. Por lo tanto es mucho más grande.

Lo que haré ahora lo tenéis que multiplicar por muchos ceros, ¿de acuerdo? Coged esto de aquí, una red de cinco puntos, cada punto es una página web. ¿Y ahora qué podemos establecer? Enlaces. ¿Que significará esto? Por ejemplo, hacer un enlace a P2 significa que des de la página web 1 hay un enlace a la página web 2. Ve hacia allí. ¿De acuerdo? Ve hacia allí. Haré lo mismo con este de aquí, el P2, haré lo mismo con el P4, el P3... y eso es todo el sistema.

Esto de aquí, en seguida veréis que tiene un nombre en matemáticas, se denomina grafo. Un grafo es la manera abstracta de representar conexiones que tiene dos elementos, los nodos que son los puntos, y las aristas, que son las flechas.

Vivimos continuamente con grafos como estos de aquí. Ahora veréis más ejemplos. Ahora tenemos esto aquí. Guardadlo en la memoria. ¿Qué haremos con este bicho de aquí? Pues lo trabajaremos. Eso que os decía. Eso es lo que se denomina grafo. Un grafo tiene dos partes: los puntos, los nodos y las aristas. Grafos típicos: pueden ser dirigidos o no dirigidos. ¿Qué significa dirigido? Que la flecha importa. Por ejemplo, cuando enlazo otra página web o per ejemplo, Twitter. Yo puedo seguir a Puyol, el jugador de futbol, pero seguramente él no me sigue a mí. O sea, no es bidireccional, sino que tú marcas la dirección. Tú sigues alguien, por lo tanto la flecha va de uno a otro.

En cambio facebook, cuando sois amigos, la flecha va en los dos sentidos, estáis ligados, sois amigos. Por lo tanto es un grafo que de hecho es no dirigido porque no hay direcciones.

Automáticamente hay un enlace entre los dos. Bien pues vamos a jugar con los grafos. Si tengo un grafo como este de aquí, ¿cómo lo podría guardar? Y ahora empieza la parte matemática. Tenéis un problema y tenéis que hacer una versión abstracta.

Empecemos con una cosa de juguete llamada TOM model, una cosa para jugar, este de aquí. Fijaos: P1 tiene enlaces a P2, P3 y P4, y P2 enlaza a las páginas P3 y P4. Si os preguntase cómo representaríais esto de aquí de manera que la información quedase grabada fácilmente, ¿cómo lo haríais? Una versión podría ser esta, un poco de andar por casa: hago una cuadrícula. Aquí pongo todas las páginas web, tengo cinco, si tuviese 2,7 billones no las podría poner, pero con cinco, puedo hacerlo.

Y aquí puedo poner las de salida y hago una crucecita. Por ejemplo, la P1 enlaza a la P2, a la P3, a la P4 y a la P5. Solo pongo una crucecita aquí: P2, P3, P4 y P5.

Aquí no pongo ninguna porque la P1 no se autoenlaza. Habitualmente una página web no se autoenlaza, no es tan interesante.Y hacemos lo mismo con la P2. Y tenemos la primera versión. Está bien, a los matemáticos nos gusta trabajar con números, y entonces, ¿qué haces? Pues lo cambias por 0 y 1.

No está mal. ¿De acuerdo? Esto es un código. 0 y 1. 1 quiere decir que hay conexión y 0 quiere decir que no hay esta conexión. Esto de aquí, ¿cómo lo guardo?

Bueno, los matemáticos tenemos una cosa que es como ir a IKEA, una especie de armarios que contienen información. Lo que en IKEA es un armario, en matemáticas es una matriz.

Las matrices son el lugar en el que guardas muchísima información. Os sorprenderíais de la cantidad de cosas que podéis poner en una matriz. Sirven para muchas cosas. De manera compacta podéis guardar muchas cosas. Yo lo que haré será coger esto de aquí y lo guardaré en una matriz. Y ahora os explicaré cómo se guardará.

¿Cómo lo guardaremos? Pues me inventaré una matriz me saco de la manga una matriz que se llama matriz de conectividad. Quiere decir que están conectados.

Por ejemplo, fijaos: P1 está conectado a P2, P3, P4 y P5. ¿Recordáis que la tabla de antes tenía 0 y 1, 1, 1, 1? Pues haré lo mismo pero lo pondré dentro de la matriz.

¿Dónde lo pondré? Aquí, en la primera fila. Esta matriz solo tiene 0 y 1. Aquí 0 quiere decir del P1 al P1 no hay conexión. Aquí 1 quiere decir que P1 envía un enlace al P2, al P3, al P4 y al P5. Y esto lo haremos con todo y tendremos una matriz.

Segundo, esta de aquí, la P2, enlaza con la P3 y la P4. ¿Qué tenemos aquí? Fijaos, todo 0 menos un islote en la P3 y la P4. Y haremos lo mismo con el P3, el P4 y el P5.

Y ya tengo una matriz. Una cosa que es real, voy a un ordenador, entro y hago clic en un enlace, la tengo de forma abstracta en una matriz.

Mirad cómo toda la información cabe aquí, en una matriz. 0 i 1. Más fácil imposible. Bien pues esta es la matriz de conectividad, me dice si hay conexión o no.

La pregunta es: esto es bueno para medir el ranking? Si yo tuviese que ordenar aquí de más popular a menos popular, ¿cuál sería la más popular? ¿Cuál sería? ¿La 1? Fijaos en una cosa... esto de aquí... Las filas quiero decir.... La primera fila indica a quién enlaza la P1.

En cambio si m e lo miro por columnas, esta de aquí es quien envía la flecha a la P1. Y solo encontramos la 5. Fijaos, a la P1, solo la apunta P5. Cuando lo miras por columnas, esto es una columna, esto me dice quién envía una flecha a la página P1. Mirad, P1 no es muy popular, solo le envía una flecha a P5, solo enlaza a P5, que es su primo. La segunda, P2. ¿Quién envía una flecha a P2? La P1 y la P4. La P1 envía flecha a P2 y el P4.... ¿Dónde está P4? 1, 2, 3,4... pues nos falta una flecha.

Pues he puesto mal flecha, porque en principio, si lo miro por columna, esto quiere decir el número de enlaces que tengo a mi página web. Muy bien, de estas, ¿cuál es la que tiene más éxito? La P4 ¿verdad? Fijaos, tiene 4. ¿Cómo lo traduzco en números?

Lo que podría hacer es sumar el número de entradas. Hagámoslo... sumo el número de entradas. Es decir, tengo aquí arriba las páginas y ¿cuántas entradas llegan? 1, 2, 2, 4 y 1. Tiene 10 flechas, tiene 10 conexiones que son todas las que hay. Reparto. ¿Quién gana? La P4. ¿Cómo lo puedo medir? Lo podría hacer de muchas maneras más complicadas... Que tiene un 4 sobre 10... si queréis puedo decir que de cada 10, 4 van a parar a la P4. O si queréis, sobre 1 podríais decir un 0,4. O en números económicos podríais decir un 40% de las entradas que hay en el sistema se va hacia la 4. Este sería vuestro David Guetta.

Si yo hiciese la ordenación, diría: la que gana es la P4. Es la más importante. Primera versión. Muy bien, si yo os dijera... esto de aquí es con cinco puntos, con cinco ordenadores.

Si yo cogiese los 2,7 billones de páginas web que hay en el mundo, ¿cómo sería la matriz? Os enseñaré un trocito muy pequeño, que es todo lo que hace referencia a la Universidad de Stanford.

No sé si veréis algo, pero es, con fondo negro... esto está en la Wikipedia. Si vais a la Wikipedia, justamente PageRank, que es como se llama este algoritmo veréis que esto de aquí es... esta es la fotografía que sale. Todo es negro y hay un punto verde allí donde hay una conexión, donde hay un enlace. Si queréis, los puntos verdes son 1 y los puntos negros quieren decir 0.

Esto de aquí es un trocito muy pequeño, para que os hagáis una idea de cuán grande es vuestra matriz.

Imaginad que la tenéis que dibujar con un lápiz, y yo haré una marquita para indicar el 0 o el 1.

Pero lo haré muy finito, de un milímetro, con un lápiz del 2. Si yo pusiese todas las páginas web del mundo en fila, son 2,7 billones, eso hace con 2,7 millones de km.

Pues ahora coged la matriz, esto es una matriz de 2,7 millones de km por 2,7 millones de km. Esto es enorme. Es monstruoso. La matriz de conectividad es una cosa muy grande. Enorme. Pero es lo que se utiliza. Bueno, sí y no. Tiene problemas.

¿Por qué tiene problemas? Ahora imaginad que cada uno de vosotros hace una página web y os enlazáis entre vosotros. Aquí podríamos decidir quién es el más popular de aquí. Pero ahora imaginad que uno tiene un enlace de David Guetta. Digo David Guetta porque me lo dicen mis hijos, pero podría ser cualquier otro. Alguien importante, un músico que os guste mucho. Y este enlaza a vuestra página web. Es lo mismo que os enlace, con todos los respetos, uno de vuestros compañeros que David Guetta? ¿Verdad que no es lo mismo? Por lo tanto no solo cuenta el peso, no cuenta por popularidad, también cuenta por autoridad.

Depende de quién te enlace, también es importante y cambia mucho. Y eso aquí no se tiene en cuenta, porque es 0 o 1. Por lo tanto, no va bien, no es un buen ranking. Este ejemplo de aquí. No es lo mismo que os enlace la IBM a la página web de vuestra empresa que os enlace el primo. Está muy bien que lo haga porque os quiere, pero IBM os puede dar de comer. Por lo tanto esto se tiene que tener en cuenta.