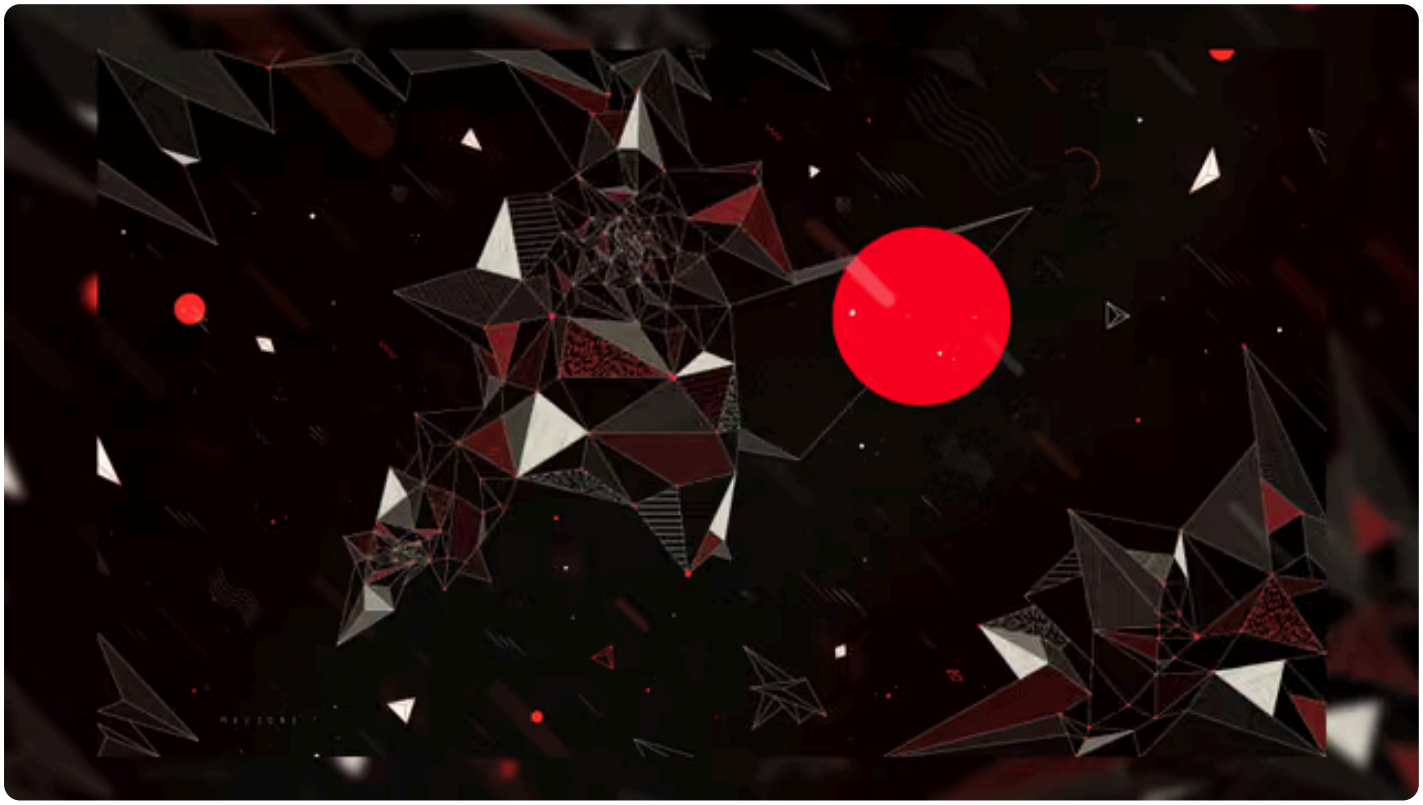


08/04/2015

Destripando Google

TXT FERNANDO SCHAPACHNIK IMG EUGENIA MALIGNE

¿Cómo organiza Google la información?

Cualquier tecnología suficientemente avanzada es indistinguible de la magia.

Arthur C. Clarke

Hay cosas un tanto mágicas que, **a fuerza de repetición, dejan de maravillarnos**. Varios vivimos en el medio de la ciudad, a kilómetros de cualquier espejo de agua y, sin embargo,

abrimos la canilla y el indispensable líquido simplemente aparece. Casi nunca pensamos en la **fenomenal obra de ingeniería** necesaria para que eso suceda y tampoco solemos recordar que no siempre fue así (y que, lamentablemente, en muchos lugares todavía no es así). Con **Internet** pasa algo similar. Con el buscador, para ser más precisos. Ponemos una frase, una palabrita, y enseguida aparecen los resultados. *Muy enseguida*. En una fracción de segundo. Casi *demasiado* enseguida, considerando la inmensa cantidad de información que hay en Internet. Entonces**, ¿qué hay detrás de la brujería que permite encontrar la página que queremos entre las casi 4,5 mil millones existentes?**

Lo que hay no es ni más ni menos que otra **obra monumental del pensamiento humano**. Un puñado ideas geniales, brillantes. Y la mayoría no se le ocurrieron a Google.

Cualquier sitio de búsqueda de Internet como Google, Yahoo o Bing, primero debe asegurarse de tener la **información disponible**. Para eso es necesario hacer dos cosas en simultáneo: **recorrer toda la web y almacenar la información que contiene**. Si vamos a reunir más de 4,5 mil millones de páginas, claramente necesitamos mucho espacio, y para eso están destinados buena parte de los **centros de cómputo** que Google tiene **por todo el mundo**, ya que la información no se guarda en un único disco rígido XXXXXXL, sino en **millones de discos rígidos normales**, distribuidos en millones de computadoras interconectadas entre sí. Eso no es nada original, aunque hay que reconocerle a Google que sí innovó sobre la forma en la que utiliza esa interconexión.

Pero lo más interesante es analizar **la parte encargada de hacer el tour por la web**, una idea sencilla que viene a ser una receta de tres pasitos nada más:

1. Empiezo por una página web (www.elgatoylacaja.com, ponele, o alguna más copada).
2. **Almaceno en una gran base de datos las palabras que aparecen**, asociadas a la página web en la que están. Por ejemplo, si en la página encuentro las palabras ‘instrucciones para armar un pollo’, voy a almacenar:
    1. instrucciones → www.elgatoylacaja.com

2. para → www.elgatoylacaja.com
3. armar → www.elgatoylacaja.com
4. un → www.elgatoylacaja.com
5. pollo → www.elgatoylacaja.com

3. Veo qué links tiene esa página y los voy siguiendo, pasando a las páginas que indican. Y es acá donde **el proceso vuelve al segundo paso**. Así, una página lleva a otras, que llevan a otras, que llevan a otras..., y de esta forma voy 'descubriendo' la famosa red.

La manera en la que Google hace esto no es obra de Google. Es muy conocida y se llama BFS: **Breadth First Search**, o búsqueda 'a lo ancho', que a su vez pertenece a un área del conocimiento humano llamada '**Teoría de grafos**'. Los grafos son una generalización de la idea de 'cosas' conectadas entre sí. O sea, **puntitos conectados con flechitas**. Esos puntitos pueden ser páginas web y las flechitas pueden ser los links; o los puntitos pueden ser ciudades y las flechitas los vuelos que las conectan; o los puntitos pueden ser vos, tu novia, tu ex, el ex de tu ex, el pibe que ahora sale con la ex de tu actual, y millones de otras cosas menos complicadas que todo ese stalkeo que hiciste. La teoría de grafos ha desarrollado muchas formas de recorrer y analizar este tipo de estructuras de redes. Y **lo mejor de todo es que este conocimiento es público**.

La idea detrás de BFS es simple, potente y, por qué no, hermosa. Si la entendemos, nos da conocimiento y nos habilita a construir otras ideas sobre ella. **En definitiva, nos da poder**. Si no la conocemos, volvemos a la magia, pero no a la magia copada, sino a la magia mala, la magia de no saber. **Esa noche oscura del pensamiento** en la que el poder para hacer cosas lo tienen otros, y nosotros no sólo no podemos, sino que tampoco entendemos.

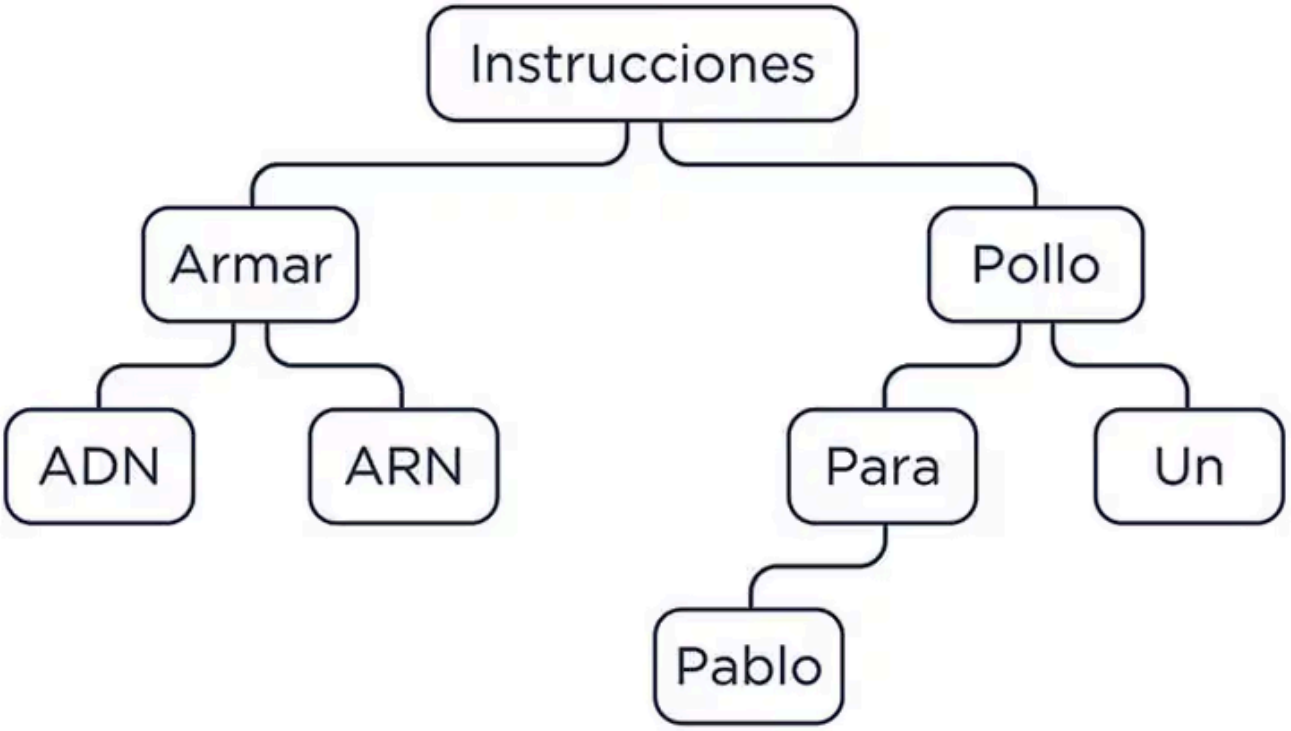
Ahora, **con la información almacenada, vamos por más**. Encontramos que la palabra pollo estaba en www.elgatoylacaja.com, y supongamos que a medida que seguimos recorriendo la red aparece en unos 75 millones de lugares más, entre ellos, la página de Wikipedia que describe al *'Gallus gallus domesticus'* (que así se llama el pobre bicho, uno

viene a enterarse). Lo cierto es que, si alguien busca la palabra 'pollo', **tal vez sería deseable que el artículo de Wikipedia quede listado (bastante) antes** que el artículo de Pablo sobre el ADN (con todo respeto por el ADN, no tanto por Pablo). Para eso Google usa, de nuevo, una idea original y otra de dominio público.

Todo muy listo el pollo, pero ¿qué sucede cuando alguien tipea P-O-L-L-O (enter) en el buscador? Tenemos ese listado inmenso de palabras, cada una de ellas con una gran cantidad de páginas web asociadas. **¿Cómo hace Google para brindarnos un resultado en menos de medio segundo?** ¿Es porque tiene computadoras poderosas y buena conectividad? En parte, pero no principalmente. ¿Es porque tiene muchas computadoras que hacen distintas partes del trabajo? En parte, pero no sólo por eso. ¿Es porque tiene un ejército de gnomos en dudosas condiciones laborales atentos a tus dudas? No, definitivamente no, pero todos lo pensamos alguna vez.

La idea responsable de la alta velocidad es también muy sencilla. Y de dominio público. Patrimonio de toda la humanidad: **los árboles**. No los árboles de la sombra, la leña y la fotosíntesis, sino otros un poco más abstractones.

El truco consiste en que **el dichoso 'listado' inmenso de palabras no se guarda como un listado**, con las palabras una a continuación de la otra, sino de la forma que los computólogos llamamos árbol:



Hay algo interesante a señalar, **una propiedad estructural del dibujo**. Si nos concentramos en 'instrucciones', vamos a ver que la palabra que está a la izquierda es anterior en sentido alfabético y la que está a la derecha es posterior. Más aún: no sólo la que está inmediatamente a la izquierda, unida por la flechita, es anterior, sino que todas las que 'cuelgan' de ella lo son. Y por la derecha pasa lo mismo, son todas mayores. Pero hay más: esa característica es estructural porque también **se da en cualquier parte del árbol.** Por ejemplo, si analizamos la palabra 'pollo', también sucede que todas las de su izquierda son anteriores y todas las de su derecha son posteriores.

Acá está la magia, y de la buena. Si las palabras estuviesen en un listado, ¿cuántos pasos nos tomaría encontrar 'pollo'? **Si el listado no tuviese orden**, dependería de dónde quedó, así que por ahí tenemos mala suerte y es la última, en nuestro ejemplo, 8 pasitos. **Si estuviese por orden alfabético**, quedaría anteúltima: 7 pasitos. Pero qué lindo el árbol que nos deja comparar la palabra que estamos buscando con la primera que aparece; aquí, 'instrucciones'. Si la que buscamos es posterior (alfabéticamente hablando), ya sabemos que va a estar a la derecha y, si fuese anterior, a la izquierda. Ahí podemos 'ir' para el lado correspondiente y repetir el procedimiento hasta encontrarla. O sea, llegamos a pollo en dos pasos. De hecho,

llegamos en a lo sumo 4 pasitos a cualquiera de las palabras que tenemos (porque, como se ve en el dibujo, en 4 pasos se acaban el árbol, el pollo y la metáfora).

¿Y si tuviese 20 millones de palabras? **Ahí viene otra propiedad de estos árboles: están *balanceados***, lo que significa que, parado en cualquier palabra, la cantidad de pasos que hay que dar para llegar a la última de ellas por izquierda, vs. la cantidad de pasos para llegar a la última por derecha, difieren en a lo sumo 1 paso. Eso, sumado a un poco más de matemática, permite demostrar que siempre vamos a encontrar la palabra que buscamos en a lo sumo una cantidad logarítmica de pasos. O sea que **encontrar una palabra entre 20 millones toma sólo 24 pasos**. Y, si incrementamos por 100 el caudal de datos (pasamos a 2000 millones), son 31 pasos. 31 pasitos te los hago con los ojos cerrados y las manos atadas detrás de la espalda, 2000 millones (como en el caso del listado), capaz demoran un toque más.

Este árbol maravilloso que acabamos de analizar se llama **árbol binario de búsqueda balanceado**, también conocido como AVL, que por suerte no es lo mismo que ABL, y que ni siquiera tiene A de Árbol o B de Balanceado o L de binario (?), sino que viene de Adelson-Velskii y Landis, los matemáticos rusos que lo inventaron. Y lo publicaron. Y es patrimonio de la humanidad.

Cuando una idea tan increíblemente potente como la del AVL, tan presente de manera casi invisible en nuestra vida cotidiana bajo diversas formas, tan abiertamente patrimonio humano, **se vuelve a la vez tan desposeída, en el sentido de que sólo la conoce una fracción muy pequeña de la población**; **estamos frente a la peor forma de la magia.** Porque a los que la conocen y la usan les da poder, y a los que la ignoran, los somete y les hace pensar que esas características que los separan de los otros son inalcanzables, insondables e incomprensibles. Porque parecen magos y no mortales que se equivocan, que tienen deseos, intenciones, restricciones, y contextos.

Aprender se trata de tomar ese capital humano que ya nos pertenecía aunque no lo supiéramos. Somos varios los que queremos que a los chicos argentinos se le enseñe Ciencias de la Computación en la escuela, para que vean, descubran y entiendan la magia buena detrás

de la tecnología; para que transformen la limitación de no saber en el poder de conocer y así puedan disfrutar de un montón de ideas que les pertenecen, aunque aún no lo sepan.

Referencias

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, y Clifford Stein. Introduction to Algorithms, segunda edición. MIT Press y McGraw-Hill, 2001. ISBN 0-262-03293-7.

Log-structured merge-tree
