La recuperación de la información en la era de la información: de las pinturas rupestres a Google*

Pablo Argibay

HISTORIA

La recuperación de la información (*information retrieval*) es el proceso de búsqueda dentro de una colección de documentos en orden a cumplir cierta necesidad particular de información (*query*). Si bien se tiende a asociar la recuperación de la información con la computación, lo cierto es que el almacenamiento de la información y su recuperación tienen una historia tan lejana como la humanidad. Tal vez las colecciones de información registrada más antiguas sean las pinturas rupestres dibujadas en las cavernas. La búsqueda de información en esta colección debería hacerse a pie recorriendo la caverna y observando cada pintura. El siguiente antecedente en guardar y recuperar información podría encontrarse en la utilización de papiros. Los romanos utilizaban etiquetas externas para ordenar y buscar determinados papiros. Posteriormente, los pergaminos unidos fueron el antecesor de los libros y la Biblioteca de Pérgamo, la segunda en importancia después de la de Alejandría, una de las colecciones más grandes de tales dispositivos. Se cuenta que los libros de la Biblioteca de Pérgamo eran de papiro y que su competidora, Alejandría, dejó de abastecer ese material y Pérgamo tuvo que copiar sus libros en un material que ellos mismos fabricaban llamado pergamino. Sin embargo, algunos historiadores aseguran que el pergamino surgió como una necesidad de crear un elemento donde escribir los libros que fuera más duradero que el papiro.

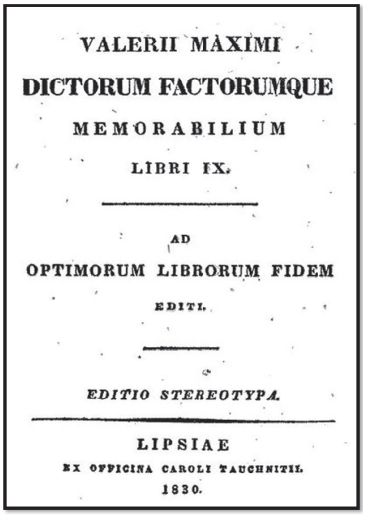
En etapas posteriores (Edad Media), sorprendentemente, la recuperación de la información se efectuaba con sistemas de búsqueda análogos a los actuales sistemas informáticos. A falta de libros y computadoras, existían individuos narradores de historias. Aparentemente, estos individuos eran detenidos en su caminar y se les preguntaba acerca de tal o cual pasaje de una historia. Imaginamos algún sistema de búsqueda de la información en la "cabeza" de estos sujetos, verdaderas bibliotecas ambulantes. Más adelante, la invención del papel y los escribas y de la imprenta (Johann Gutenberg, 1450) hicieron que la información estuviera guardada en libros, al cuidado de monjes en un principio, y luego en bibliotecas. Imaginamos diversas señales y sistemas de clasificación por las cuales era eficaz la búsqueda de determinada información específica. Las clasificaciones eran entonces señales tan fundamentales para motorizar una búsqueda, como los libros mismos. Aparentemente el primer sistema jerárquico de almacenamiento de la información fue el *Factorum ac dictorum memorabilium libri IX* (nueve libros de hechos y dichos memorables) (Valerius Maximus, 30 d.C.) (Figura 1).

La complejidad de las modernas bibliotecas obligó a la creación de sistemas de guardado y recuperación de las informaciones más sofisticadas:

- Sistema decimal de Dewey (1872)
- Catálogo de tarjetas (1900)
- Microfilms (1930)
- Catálogo de lectura por máquina (MARC) (1960)

Las búsquedas seguían siendo jerárquicas y necesitaban de amplios conocimientos en dichos sistemas de clasificación, por lo que quedaban limitadas a bibliotecarios y bibliotecarias. Con el advenimiento de la computación (1940-1950), aparecieron sistemas automatizados de búsqueda como el SMART de Cornell (1960).

Figura 1. Versión de 1830 de los *Factorum ac dictorum*. Fuente: http://en.wikipedia.org/wiki/Factorum_ac_dictorum_memorabilium_libri_IX


VALERII MAXIMI
DICTORUM FACTORUMQUE
MEMORABILIUM
LIBRI IX.
AD
OPTIMORUM LIBRORUM FIDEM
EDITI.
EDITIO STEREOTYPA.
LIPSIAE
EX OFFICINA CAROLI TAUCHNITII.
1830.

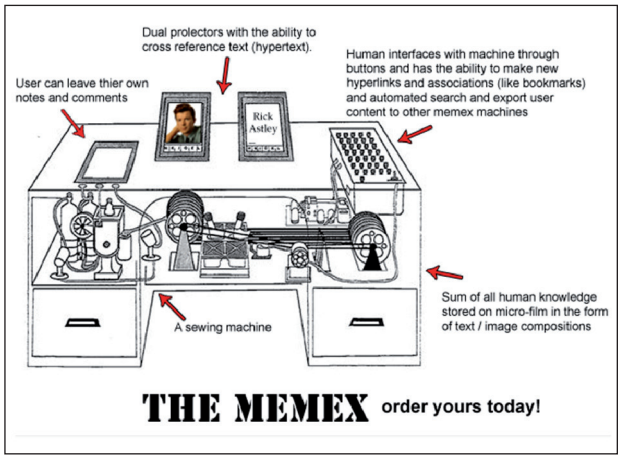
* El presente trabajo es de naturaleza monográfica no original y está basado fundamentalmente en una síntesis con modificaciones de Langville y Meyer, 2006.[3]

Laboratorio de aprendizaje biológico y artificial, ICBME-IU. Hospital Italiano de Buenos Aires
Correspondencia: pablo.argibay@hospitalitaliano.org.ar

El eslabón perdido de la Web: el Memex

En 1945 Vannevar Bush, ingeniero y científico (Director de la oficina de investigación y desarrollo científico), escribió un ensayo "As We May Think".[1] En ese artículo Bush describe el Memex, un dispositivo electromecánico de sorprendente parecido conceptual con las computadoras modernas y la Web. Memex trataba de utilizar los supuestos mecanismos humanos de asociación de ideas, para asociar documentos y colección de documentos relacionados (Figura 2).

Figura 2. Esquema del Memex aparecido en la revista LIFE. Fuente: http://www.internaldrive.com/wp-content/uploads/2009/03/memex.png



La World Wide Web (Tim Berners-Lee, 1989)

Se ha dicho que la revolución WWW inicia el dominio total de la era de la información por sobre la era industrial. Sin embargo, las cantidades desproporcionadas de información ubicadas en la "Web" trajeron un problema crítico para la recuperación de la información: ¿cómo conseguir un sistema de búsqueda adecuado para recuperar la información específica que una persona necesitaba?

En 1998, el "análisis de conexiones" (*link analysis*) produjo una nueva revolución en la recuperación orientada de la información. Los más exitosos "motores de búsqueda" comenzaron a utilizar esta técnica.

Clasificación *(ranking)* de las páginas web por popularidad

Los modelos de búsqueda que explotan la estructura de hiperconectividad de la Web se llaman "modelos de análisis de conexiones". En 1998, en la Universidad de Stanford, dos doctorandos en Ciencias de la computación (Sergey Brin y Larry Page) estaban trabajando en sus dormitorios en un proyecto (*PageRank*), que devendría en el gran éxito de Google. Finalmente, en una presentación pública en el 7° Congreso Internacional de la Word Wide Web, presentaron su *paper*, base del *PageRank*.[2]

Ideas subyacentes en el *PageRank*. En primer lugar es necesario entender a la Web como un grafo.* Los nodos representan a las páginas web y las conexiones, a los hipervínculos (*hyperlinks*). Lo que llega a un nodo es una conexión de entrada y lo que sale, una conexión de salida (Figuras 3 y 4).

Figura 3. Grafo que representa una "web" de 6 páginas y sus conexiones.

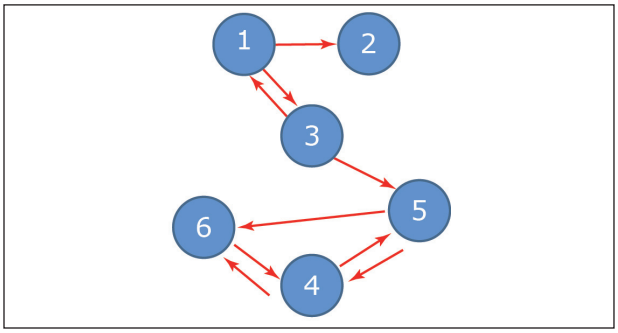
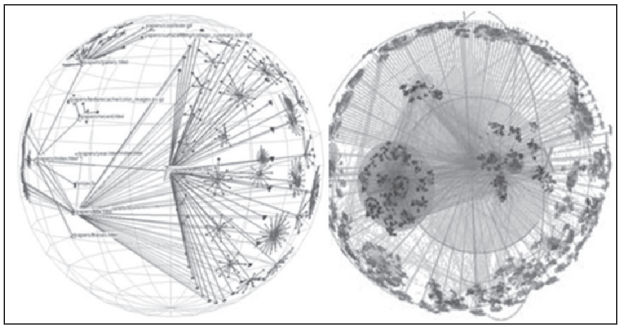


Figura 4. Representación de subconjuntos conectados en la Web a través de un mapa de Munzner y Hyun. El mapa de la izquierda es un espacio tridimensional hiperbólico. El mapa de la derecha representa una web de 535.102 nodos y 601.678 conexiones. (Atlas of Cyberspace 2001)



Las conexiones se pueden metaforizar como recomendaciones. Una conexión de A a B es una especie de recomendación o aprobación de A para B. Una página con más recomendaciones sería más importante que una página con pocas conexiones de entrada. También el estatus del que recomienda es importante. Por otra parte, la baja discriminación de los que recomiendan disminuye el peso del recomendado. En síntesis, la tesis subyacente en el *PageRank* es: una página web es importante si es recomendada por otras páginas importantes. Parece trivial pero, desde el punto de vista matemático, la complejidad de esta red de recomendaciones y *rankings* puede ser formalizada en una ecuación matemática simple y bella.

* Un grafo es un conjunto de puntos (vértices) en el espacio, que están conectados por un conjunto de líneas (aristas) www.matediscreta.8k.com/grafos.htm. Otros conceptos básicos son: Dos vértices son adyacentes si comparten la misma arista. Los extremos de una arista son los vértices que comparte dicha arista. Un grafo se dice que es finito si su número de vértices es finito.

La matemática de *PageRank*

Brin y Page son los inventores de *Page Rank** y fundadores de Google Inc., el 4 de septiembre de 1998. Este motor de búsqueda superó al otro más popular de la época, AltaVista, que había sido creado en 1995. Brin y Page conceptualmente diseñaron el *ranking* de *Page Rank* con una sumatoria (ecuación 1), basada en los índices bibliométricos utilizados en la estructura de citación de trabajos científicos:

Ecuación 1

$$r(Pi) = \sum_{PjeB_{Pi}} \frac{r(Pj)}{|Pj|}$$

B_{Pi} = conjunto de páginas que apuntan a Pi.
|Pj| = número de salidas desde Pj.

El ranking *(PageRank)* de una página Pi (r(Pi)) es la suma de los *rankings* de todas las páginas que van hacia Pi.

El problema con la ecuación 1 es que r(Pj) es desconocido. Para solucionarlo se diseñó un procedimiento iterativo, asumiendo que al inicio todas las páginas tendrían un *PageRank* igual (1/n; donde "n" es el número de páginas en el índice Google de la Web). La ecuación 1 se aplica para computar r(Pi) para cada página Pi en el índice. En el procedimiento iterativo:

Ecuación 2

$$rk + 1_{(Pi)} = \sum_{PjeB_{Pi}} \frac{rk(Pj)}{|Pj|}$$

$r_{k+1}(Pi)$ = *PageRank* de Pi en la iteración k+1

El proceso se inicia con $r_0(Pi) = 1/n$ para todas las páginas Pi y se repite hasta una eventual convergencia hacia un valor estable. Por ejemplo para la figura 3, después de pocas iteraciones obtenemos los valores expresados en la tabla 1. Las ecuaciones 1 y 2 computan el *ranking* una página a la vez. El uso computacional de matrices, obviamente, reemplaza "Σ" y en cada iteración computa un vector del *PageRank*. Este usa un vector de 1 × n para mantener los valores del *PageRank* para todas las páginas en el índice. La implementación algebraica requiere una matriz "H" de n × n y un vector fila π^T de 1 × n. La matriz H es una matriz normalizada por fila con $H_{ij} = 1/|P_i|$ si existe una conexión entre i a j, o 0 de cualquier otra manera. Tomando como ejemplo simple el grafo de la figura 1, tenemos una matriz H = donde los elementos no 0s son probabilidades.

Los elementos no 0s de las filas i son las salidas de las páginas i y los elementos no 0s de las columnas i corresponden a las entradas de la página i. La matriz H para el grafo de la figura 3 es la siguiente (tomado de A.N. Langville y C. Meyer, 2006[3]):

Tabla 1. Aplicación de la Ecuación 2 (luego de algunas iteraciones) a la figura 3. Tomado de A.N. Langville y C. Meyer, 2006.

Iteración 0	Iteración 1	Iteración 2	Ranking
r0(P1) = 1/6	r1(P1) = 1/18	r2(P1) = 1/36	5
r0(P2) = 1/6	r1(P2) = 5/36	r2(P2) = 1/18	4
r0(P3) = 1/6	r1(P3) = 1/12	r2(P3) = 1/36	5
r0(P4) = 1/6	r1(P4) = 1/4	r2(P4) = 17/72	1
r0(P5) = 1/6	r1(P5) = 5/36	r2(P5) = 11/72	3
r0(P6) = 1/6	r1(P6) = 1/6	r2(P6) = 14/72	2

	P1	P2	P3	P4	P5	P6
P1	0	1/2	1/2	0	0	0
P2	0	0	0	0	0	0
P3	1/3	1/3	0	0	1/3	0
P4	0	0	0	0	1/2	1/2
P5	0	0	0	1/2	0	1/2
P6	0	0	0	1	0	0

Introduciendo el vector $\pi^{(k)T}$, el vector de *PageRank* en la iteración k^{th}, escribimos la ecuación 2 como:

Ecuación 3

$$\pi(k+1)\ T = \pi(k)T\ \mathrm{H}$$

El *ranking* de Page y Google tienen algo de fenómeno "Eureka"** en el imaginario popular, lo mismo que *Facebook*. Sin embargo, son fruto de una gran idea y un intenso trabajo computacional y matemático. *PageRank* puede resolverse como un sistema de ecuaciones lineales y de hecho es un ejemplo de enseñanza en cursos avanzados de álgebra lineal y métodos numéricos.

SÍNTESIS

El hombre desde sus inicios necesita información, para tomar decisiones, para resolver problemas o simplemente por ese fenómeno no bien explicado científicamente de la "curiosidad". Históricamente ha ido desarrollando sistemas para almacenar, relacionar y recuperar esa información. Las computadoras han cambiado el mundo y lo han cambiado ya que están llevando a términos jamás imaginados la solución a aquel problema de la necesidad de información. En definitiva, en su aspecto más esencial las computadoras son

* La patente la tramitó Larry Page en 1998 (US Patent # 6285999).

** Durante el proceso de solucionar un problema determinado, se llama fenómeno "Eureka" al descubrimiento o creación de una idea o dispositivo original, o a la resolución espontanea del problema a partir de un hecho fortuito y aparentemente no relacionado con el problema en estudio. Se denomina "Eureka" (¡lo he hallado!): en general se da este nombre al fenómeno con el cual se describe la espontaneidad de un hecho creativo (descubrimiento, solución a un problema), a partir de hechos aparentemente no relacionados con el problema en sí. Se denomina así a partir del famoso aunque no comprobado episodio de Arquímedes en su bañera. Un hecho similar sería el de "la caída de la manzana a Newton".

máquinas con una capacidad tremenda de almacenar, relacionar y brindar información. Internet ha multiplicado por factores inmensos dicha capacidad, hoy es una gigantesca trama de información. Los creadores de las maquinarias de búsqueda y diseñadores de *ranking* han intentado hacer la búsqueda de la información más dinámica, autoorganizada y superconectada. La tarea del futuro será hacerla fiable. La evidencia, siempre la evidencia, la búsqueda de la mejor y más fiable información para tomar una decisión sobre hechos con alto grado de incertidumbre.

REFERENCIAS

1. Vannevar Bush. As we may think. Atlantic Monthly, 176(1):101-108,1945. Citado en Langeville y Meyer 2006.

2. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 33:107-17,1998.

3. Langville A.N., Meyer C. Google's Page Rank and beyond: the science of search engine rankings: New Jersey: Princeton University; 2006.