Desde Memex a la Web Semántica

Daniel Gayo Avello

Última modificación: Mon, 15 May 2023 09:36:04 GMT

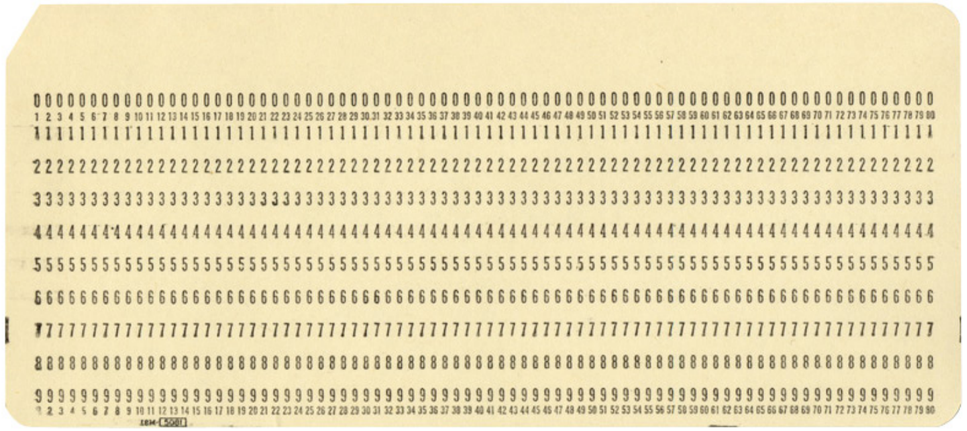
Aviso para navegantes

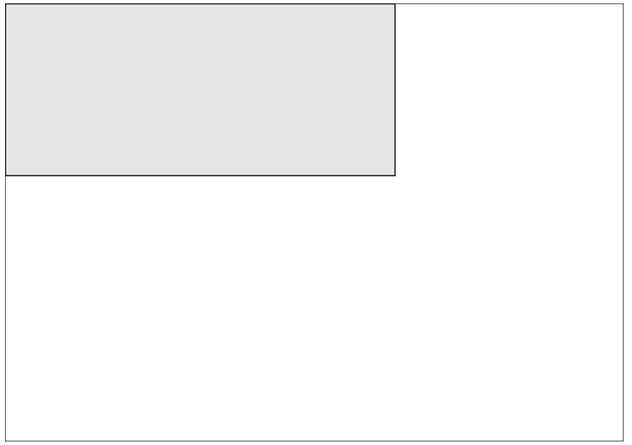
- Las clases de teoría no van a estar organizadas en lecciones claramente definidas.
- Se realizará un viaje desde los años 1950 hasta el presente estudiando cómo se almacena, procesa, difunde y busca información.
- En ese viaje nos encontraremos primero con Internet, más tarde la Web, después los directorios y sólo al final los motores de búsqueda.
- Se verán otras propuestas para solucionar la sobrecarga de información: filtrado colaborativo, recomendación de contenidos y Web Semántica.
- Veremos también cómo la Web está sesgada y es un entorno adversarial a la vez que efímero...
- ...pero con un potencial inmenso para tratar de minar conocimiento latente en ella.
- Una vez hecho el recorrido histórico volveremos atrás para ver con más detalle aspectos concretos.

Almacenamiento y tratamiento automatizado de información

- Los primeros ordenadores electrónicos eran poco más que calculadoras y no requerían dispositivos de almacenamiento externo sofisticados.
- Aún así, disponían de métodos para conservar código y datos y no tener que introducirlos "manualmente" con cada ejecución.
- Cintas y tarjetas perforadas fueron muy utilizadas (sobre todo las segundas) durante los 1950s y 1960s.
- El formato de tarjeta de IBM permitía almacenar en "modo texto" 80 caracteres (harían falta 4 tarjetas para codificar un tuit en inglés).
- Un bloque de tarjetas de una pulgada (2,54 cm) contenía 143 tarjetas y almacenaría unas 1.800 palabras.
- Así, una comunicación a congreso "típica" (5000 palabras) ocuparía 18,73 x 8,26 x 7,06 cm, es decir, algo más de 1000 cc.

Almacenamiento y tratamiento automatizado de información





Almacenamiento y tratamiento automatizado de información

- Era posible almacenar grandes cantidades de texto en tarjetas perforadas pero resultaba incómodo.

- La Seguridad Social de los EE.UU. mantenía toda la información sobre los trabajadores del país en tarjetas perforadas y presionó a IBM para solucionar esta situación.
- A mediados de 1950s IBM presentó el estándar *de facto* para almacenamiento en cinta magnética (el IBM 726).

Almacenamiento y tratamiento automatizado de información

- Las cintas magnéticas ocupaban mucho menos espacio, eran muchísimo más rápidas de procesar (7500 caracteres por segundo frente a 133 con las tarjetas) y podían re-escribirse.
- El gran problema de las cintas magnéticas era su acceso secuencial
- En 1956 IBM presentó el 305 RAMAC que incluía una unidad de almacenamiento en disco magnético, el IBM 350, con acceso aleatorio y capacidad para 5 millones de caracteres (62.500 tarjetas perforadas, 2 rollos de cinta magnética, o 750.000 palabras).
- El desarrollo de cintas y discos magnéticos continuó, aumentando densidad de almacenamiento y velocidad de acceso.
- Las tarjetas perforadas quedaron relegadas a introducción de datos hasta su desaparición a mediados de 1970s.
- En aquella época un rollo de cinta magnética podía almacenar 180MB y la unidad de disco IBM 3340, 70MB.

Almacenamiento y tratamiento automatizado de información

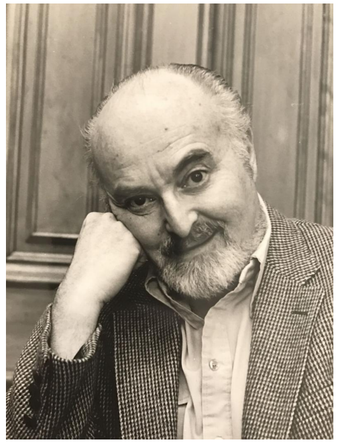
- El período 1950s-1970s estuvo caracterizado por el uso de *mainframes* para almacenar y procesar registros y transacciones, esto es, información estructurada.
- Existía también una cantidad enorme de información textual con poca o ninguna estructura que crecía de un modo continuo y debía ser consultada con frecuencia (patentes, jurisprudencia, informes técnicos, memorandos, etc.)
- A mediados de los 1950s y en 1960s se implementaron múltiples sistemas de búsqueda en distintas organizaciones.
- Solo en EE.UU. existían en 1966 más de 150 sistemas automatizados para consultar información textual.
- Algunos de los trabajos más influyentes en el campo del tratamiento y recuperación de información surgieron en esta época (véase Stevens, 1970).
- Veamos algunos hitos fundamentales...

Almacenamiento y tratamiento automatizado de información

- Hans Peter Luhn (IBM) es el pionero del área.
- En 1957 describió un método estadístico para codificar y recuperar información textual de forma totalmente automática.
- En 1958 describió una técnica para obtener resúmenes extractivos automáticamente.
- Luhn proponía utilizar la frecuencia de aparición en el texto de las distintas palabras, obviando las poco frecuentes y las demasiado comunes, introduciendo así el uso de la frecuencia de los términos en cada documento y las listas de *stop words* (palabras vacías). Ambas técnicas aún siguen vigentes.

Almacenamiento y tratamiento automatizado de información

- En 1960 Melvin E. Maron y J.L. Kuhns propusieron una alternativa aritmética a la búsqueda booleana (los términos de la consulta están o no presentes en los documentos).
- Esto permitiría calcular para cada documento una cifra que indicase su mayor o menor grado de relevancia en relación con la consulta planteada y mostrar los resultados de una consulta como una lista ordenada por relevancia decreciente.
- Ellos son los primeros en señalar que la ponderación de los distintos términos tanto en la consulta como en los documentos de la colección es fundamental.



Almacenamiento y tratamiento automatizado de información

- Desde **mediados de los 1960** (y hasta bien entrados los 1990) Gerard Salton y su equipo (cabe destacar a Christopher Buckley, Michael Lesk, Mandar Mitra o Amit Singhal) desarrollaron el **sistema de recuperación de información** SMART introduciendo toda una serie de conceptos de gran influencia posterior:
  - el modelo vectorial de documentos,
  - la utilización de la función coseno para comparar consultas con documentos [1],
  - algoritmos de *stemming*
  - o el uso de diccionarios de sinónimos y co-ocurrencias.

Almacenamiento y tratamiento automatizado de información

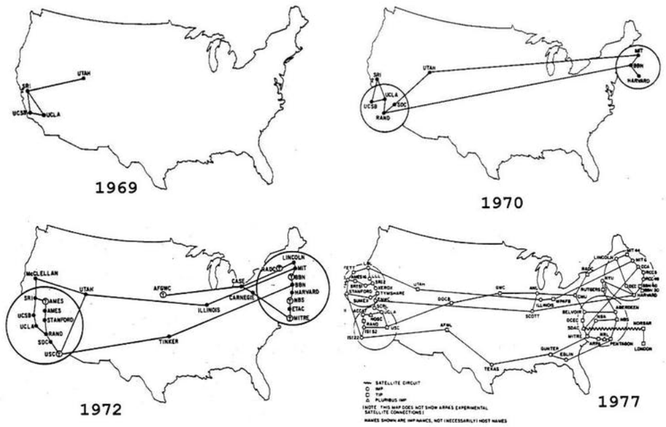
- En 1972 Karen Spärck-Jones introdujo la idea de que un término no sólo es relevante si aparece frecuentemente en un texto sino que es más valioso cuanto más raro, esto es, cuanto menor es el número de documentos de la colección en que aparece.
- Esto es lo que se conoce como *idf* que, al combinarse con la frecuencia de aparición de los términos (Luhn, 1957), ha dado lugar a uno de las formas de ponderación de términos más conocidas y utilizadas, *tf*idf*, según la cual **la relevancia de un término es directamente proporcional a la frecuencia de aparición en un documento e inversamente proporcional al número de documentos en que aparece**.
- En 1976 Stephen E. Robertson y Spärck-Jones sentaron las bases del **modelo probabilista de recuperación de información**.

Almacenamiento y tratamiento automatizado de información

- A finales de los 1970s, **tras dos décadas de investigación**, el campo contaba con unas bases teóricas sólidas que ofrecían diversas técnicas para el desarrollo de sistemas de recuperación de información con un rendimiento adecuado.
- Hasta entonces todos estos sistemas se habían diseñado y evaluado con **colecciones de documentos homogéneas y relativamente pequeñas**.
- Sin embargo, **eso iba a cambiar...**

Internet y la sobrecarga de información

- En 1969 comenzó a operar la red *ARPANET* que evolucionaría en los años 80 hasta convertirse en lo que hoy conocemos como Internet.
- Internet al acomodar otras redes de intercambio de información (como *USENET*) y ofrecer soporte para la Web (que ha dado acceso a la práctica totalidad de servicios integrados en Internet) ha contribuido en enorme medida (junto con el tremendo abaratamiento del soporte en disco y la explosión de los sitios con contenidos generados por usuarios) a la explosión de información textual que se vive hoy en día.



Internet y la sobrecarga de información

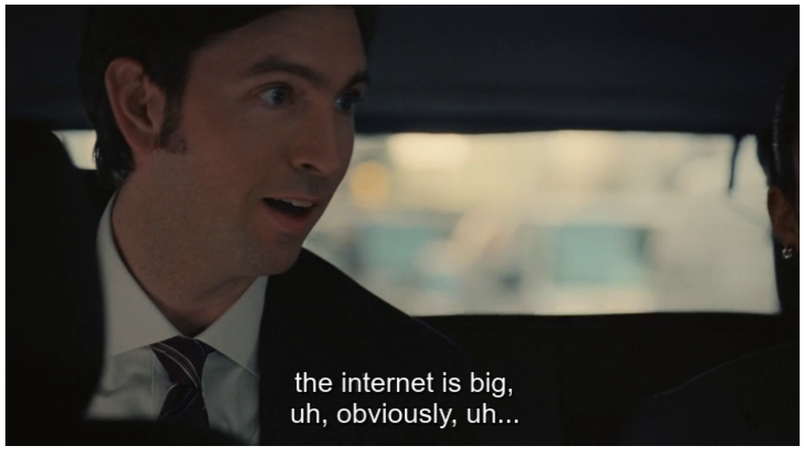
- Maron y Kuhns (p. 217) afirmaban en 1960 que se estaban generando documentos a ritmos alarmantes. Cinco décadas después no tenemos superlativos que den idea de la tasa de producción de información que ha alcanzado la humanidad.
- Bertram Myron acuñó en 1964 el término ***information overload*** (p. 857, p. 858) y ya se remontaba a Vannevar Bush para contextualizar el problema de que un exceso de información dificultara la toma de decisiones.
- En 1971 Herbert Simon advertía de la consecuencia más obvia de la sobrecarga de información: agotar la atención del público...

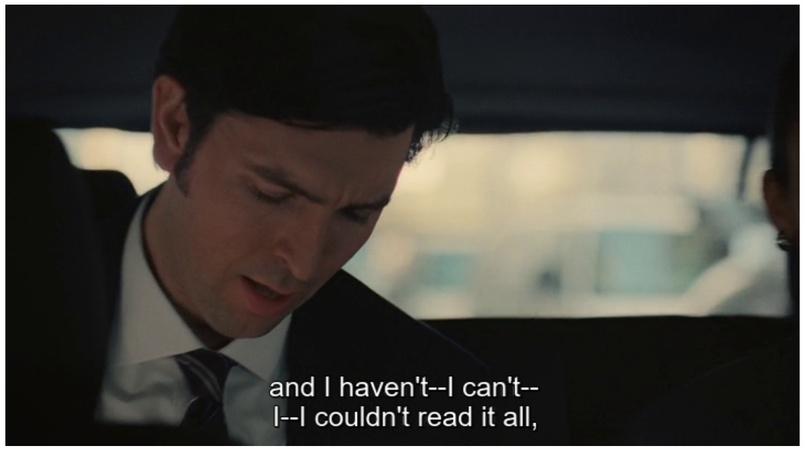
There are a number of obvious difficulties associated with the so-called "library problem" (i.e., the problem of information search and retrieval). The one usually cited relates to the fact that documentary data are being generated at an alarming rate (the growth rate is exponential—doubling every 12 years for some libraries), and consequently considerations of volume alone make the problem appear frightening. However, the heart of the problem does not concern size, but rather it concerns meaning. That is to say, there have been a number of "hardware" solutions to the problem of library size (e.g., use of microfilm, microcards, minicards, magnacards, etc.), but the major difficulties associated with the library problem remain, namely, the identification of content, the problem of determining which of two items of data is "closer" in meaning to a third item, the problem of determining whether or not (or to what degree) some document is *relevant* to a given request, etc.

The obverse of a population problem is a scarcity problem, hence a resource-allocation problem. There is only so much lettuce to go around, and it will have to be allocated somehow among the rabbits. Similarly, in an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Internet y la sobrecarga de información

- Algunas cifras (recientes pero seguramente ya obsoletas **hoy**):
  - La web superficial contiene 5.630 millones de páginas web.
  - En 2001 se estimaba que la web profunda (no indexable) podía ser 400 o 550 veces mayor que la web superficial.
  - Reuters publica 2 millones de historias al año.
  - Sólo en 2018 Elsevier publicó 500.000 artículos científicos.
  - En un segundo se publican 9.120 tuits, se hacen 85.480 consultas en Google o se envían 2,9 millones de correos.





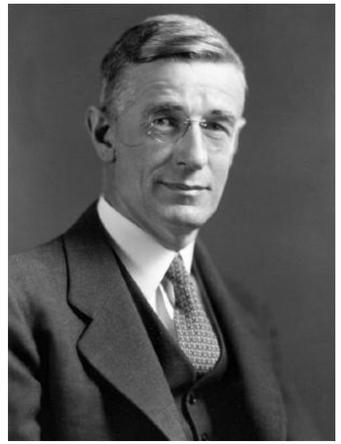
Internet y la sobrecarga de información

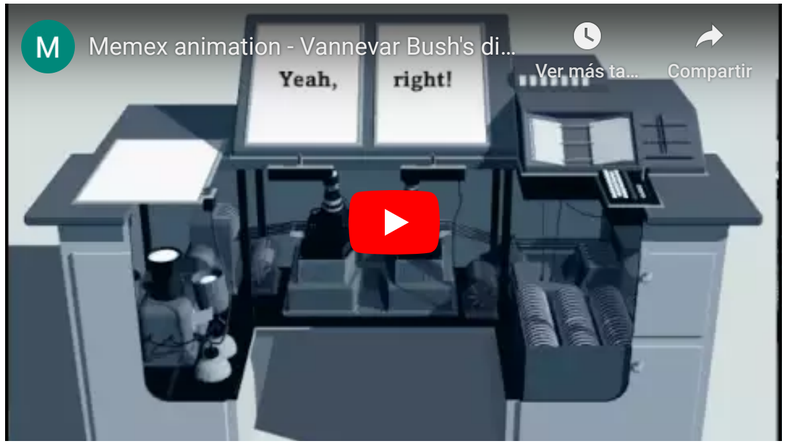
"La biblioteca total" (Borges)

- La Web se asemeja a la biblioteca descrita por Borges.
- Su tamaño es desconocido, no todos los contenidos son realmente interesantes, ni siquiera tienen por qué ser veraces.
- Los problemas de sesgos y veracidad se abordarán llegado el momento, ahora nos interesa más cómo se ha venido afrontando el problema de la sobrecarga de información en la Web.
- Pero antes, **¿cómo surgió la Web?**

La Web antes de la Web

- En 1945 Vannevar Bush describió un dispositivo denominado *memex* que serviría como un complemento a la memoria humana almacenando libros, artículos, comunicaciones personales y que, además, permitiría crear enlaces entre distintos elementos para construir nuevo conocimiento.





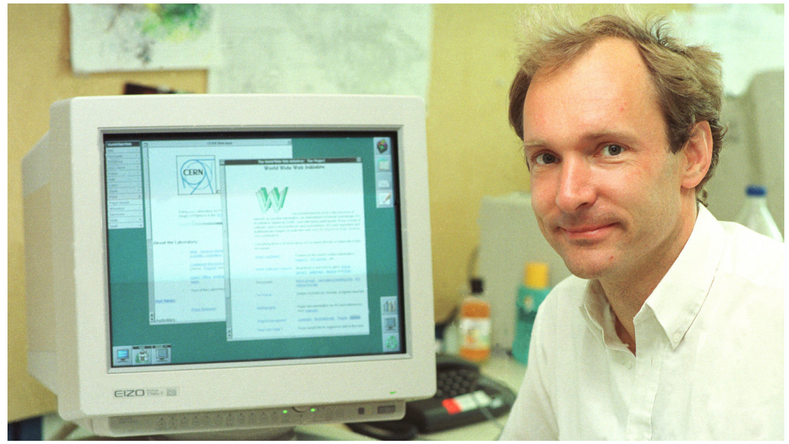
- Los conceptos asociados a *memex* influyeron enormemente tanto en el desarrollo del hipertexto, como en el diseño de los primeros interfaces gráficos de usuario.

La Web antes de la Web

- El hipertexto es un *"conjunto estructurado de textos, gráficos, etc., unidos entre sí por enlaces y conexiones lógicas"*.
- El concepto de hipertexto fue acuñado por Ted Nelson en 1963-1965, sin embargo, fue presentado al gran público por Douglas Engelbart en 1968 en la *"Madre de todas las demos"*.
- Entre los 1960s y los 1980s se propusieron y desarrollaron múltiples sistemas hipertextuales:
  - Xanadu (*vaporware*).
  - NLS.
  - ZOG.
  - KMS.
  - NoteCards.
  - Intermedia.
  - Symbolics Document Examiner.
  - HyperCard (ejemplos).
- En muchos sentidos estos sistemas anticiparon aspectos de la Web pero hubo uno crucial que ninguno supo ver: **crear un hipertexto distribuido a través de Internet.**

La Web como sistema de recuperación de información

- En 1989 Tim Berners-Lee propuso el desarrollo de la Web en el CERN como un medio para evitar la pérdida de información, inevitable en una organización de gran tamaño, y facilitar el acceso a la información disponible (bases de datos, directorios telefónicos, etc.)
- Dos características de la propuesta original permitieron transformarla en la Web actual:
    1. su **naturaleza distribuida** (los documentos pueden residir en máquinas distintas)
    2. la posibilidad de establecer vínculos (**enlaces**) entre documentos.
- Por otro lado, Berners-Lee insistía en la necesidad de construir un sistema que estimulase a los usuarios a incorporar nueva información, haciéndolo así aun más útil y atractivo, de tal forma que el conjunto de documentos creciese de forma continua. (Fenómeno estigmérgico).



La Web como sistema de recuperación de información

- Berners-Lee reflexiona sobre posibles problemas y métodos para recuperar información en un sistema como el que proponía.
  - Alerta sobre los inconvenientes de utilizar palabras clave para localizar documentos.
    Por ejemplo, ¿qué problemas plantean la polisemia y la sinonimia?
  - Sugiere la posibilidad de establecer enlaces, no sólo con documentos, sino también con conceptos facilitando la existencia de enlaces "indirectos" entre documentos de temática similar.
    Véase, p.ej. schema.org
- Así pues, la propuesta original planteaba construir la Web sobre una base semántica más o menos sólida, partiendo de nodos conceptuales enlazados desde los distintos documentos.
- Por otra parte, para explotar las ventajas de los enlaces indirectos antes mencionados, los enlaces entre documentos y conceptos deberían ser bidireccionales (ver *A Short History of Bi-Directional Links*).
- Lamentablemente, para simplificar la creación de enlaces, estos fueron unidireccionales y en ninguna de las versiones de HTML se incluyó nada similar a los "nodos conceptuales".
- Las primeras versiones del lenguaje HTML permitían dar título a los documentos, formatear texto (párrafos, listas, cabeceras, etc.) y crear enlaces a otros recursos.
- Como consecuencia, la Web se convirtió en un artefacto diseñado para crecer de un modo cada vez más acelerado sin proporcionar ningún mecanismo para localizar información.

Los primeros directorios y motores de búsqueda

- El primer servidor web (`info.cern.ch`, antes `nxoc01.cern.ch`) entró en funcionamiento en 1990.
- A finales de 1992 existían alrededor de 20.
- En esas circunstancias era fácil mantener de manera manual un directorio de sitios web y organizaciones como el CERN o el NCSA gestionaban índices a los que iban añadiendo las notificaciones de nuevos sitios que recibían por correo electrónico.
- Sin embargo, a finales de 1993 había ya 623 servidores web y su número se duplicaba cada 3 meses, existiendo en diciembre de 1994 más de 10.000 (Gray, 1995).
- Esto hacía muy difícil (aunque no imposible) mantener manualmente un índice de sitios web, dejando a un lado el hecho de que muchos webmasters no notificaban su existencia a los directorios existentes.
- La carencia de un sistema adecuado para poder localizar los distintos servidores y documentos en la incipiente Web se había convertido ya en un gran problema.

Los primeros directorios y motores de búsqueda

- Tan sólo en *WWW94* se presentaron nueve trabajos relativos al indexado automático de documentos y a la búsqueda de información.
- Cabe destacar los sistemas creados por Martijn Koster (*ALIWEB*) y Oliver A. McBryan (*WWW Worm*).
- Ambos desarrollaron programas para explorar la Web de manera automática, saltando de enlace en enlace y almacenando información sobre las páginas visitadas en una base de datos para su posterior consulta por parte de los usuarios.
- *ALIWEB* comenzaba su exploración a partir de sitios web registrados manualmente por sus administradores almacenando una información relativamente escasa para cada documento indexado (título, descripción y algunas palabras clave) lo que limitaba las posibilidades de los usuarios al realizar sus consultas.
- En el caso de *WWW Worm* no queda muy claro cómo se construía la base inicial de sitios web para realizar el indexado, la información almacenada era aún más parca (título del documento y textos utilizados en los enlaces que apuntan al mismo) y se consultaba (internamente) mediante la orden UNIX `egrep`.

Los primeros directorios y motores de búsqueda

- Otros sistemas destacables, similares a los anteriores y desarrollados en la misma época fueron *Jumpstation*, *Wanderer*, *WebCrawler* y *Lycos*.
- *Jumpstation*, implementado por Jonathon Fletcher (1994), fue uno de los primeros motores de búsqueda. Entró en funcionamiento en diciembre de 1993 y lo hizo de manera errática hasta quedar desatendido en abril de 1994.
- *Wanderer* fue inicialmente desarrollado para descubrir nuevos sitios web, posteriormente se utilizó para medir la expansión de la Web entre junio de 1993 y junio de 1995 (Gray, 1995) y, finalmente, para construir el buscador *Wandex* (*Wanderer Index*).
- *WebCrawler* (Pinkerton, 1994) supuso una mejora respecto a *ALIWEB* o *WWW Worm* puesto que indexaba todo el texto de las páginas que exploraba. Esta estrategia permitía ofrecer más documentos para las consultas de los usuarios pero, al aumentar el número de páginas indexadas, reducía de manera drástica la **precisión** de las respuestas.
- *Lycos* (Mauldin y Leavitt, 1994) constituyó una iniciativa intermedia entre *ALIWEB* y *WebCrawler* puesto que no indexaba el texto completo de los documentos ni únicamente su título y descripción. En su lugar generaba una versión "ligera" constituida por el título, las veinte primeras líneas y las cien palabras más relevantes.
- También podéis leer algo más sobre otros buscadores de la época como *Inktomi* o *Altavista*.

Para saber más: *It search-engines when it's search-engine-time*.

Los primeros directorios y motores de búsqueda

- Los buscadores podían enfrentarse al enorme crecimiento de la Web en mejores condiciones que los índices construidos de forma manual.
- Sin embargo, estos últimos ofrecían otras ventajas (organización en categorías jerárquicas, posible revisión por parte de "editores" especializados, referencias cruzadas, etc.) que eran valoradas por los usuarios.
- Por ejemplo, *Galaxy*, *Yahoo!* y *ODP* se construyeron siguiendo esta línea.

Los primeros directorios y motores de búsqueda

Galaxy

- *Galaxy* empezó a dar servicio a comienzos de 1994.
- Según el anuncio original se habían empleado métodos semiautomáticos para construir la base de datos original (que incluía no sólo páginas web sino también servidores Gopher y WAIS).
- *Galaxy*, como sería común en todos los directorios posteriores, permitía que los administradores de sitios web notificasen la dirección de su sitio para su inclusión en una categoría previamente seleccionada entre las disponibles en la jerarquía, posteriormente, un editor revisaba el sitio web y decidía acerca de su inclusión en el directorio.

Los primeros directorios y motores de búsqueda

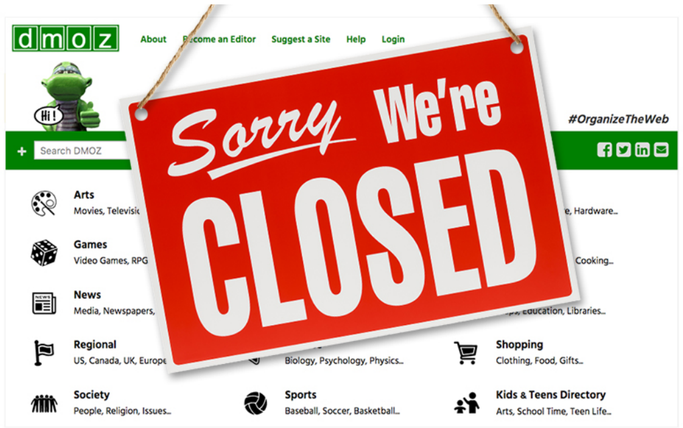
Yahoo!

- El sitio web precursor de *Yahoo!*, *Jerry's Guide to the World Wide Web* fue creado a comienzos de 1994 como un proyecto personal y se transformó en un directorio comercial en 1995.
- Al igual que *Galaxy*, disponía de categorías predefinidas en las que los administradores de sitios web pueden solicitar la inclusión que, también, es revisada por empleados de la empresa.
- *Yahoo!* retiró el directorio el 31 de diciembre de 2014.

Los primeros directorios y motores de búsqueda

ODP

- *ODP*, antes *DMoz*, antes *NewHoo* y, aún antes, *GnuHoo*, fue fundado en 1998.
- Su estructura y funcionamiento es similar a la de *Galaxy* o *Yahoo!* con la diferencia de que los editores no formaban parte de la plantilla de la empresa sino que realizaban su labor de manera desinteresada.
- El directorio fue clausurado el 17 de marzo de 2017 y con él terminó el último intento de organizar la información en la Web con editores humanos.



Estado de la Web pre-1998

- Existía toda una serie de recursos para la búsqueda de información en la Web que, sin embargo, seguían siendo insuficientes y no especialmente adecuados.
- Por un lado, no parecía que los directorios al utilizar editores humanos, pudiesen organizar todos los sitios web existentes (mucho menos todas las páginas) al mismo ritmo que aparecían.
- Por otro lado, aunque había quien afirmaba que era posible escalar los buscadores e indexar la Web al mismo ritmo que crecía, otros discrepaban.
- Steve Lawrence y C. Lee Giles (1998) analizaron la "cobertura" de distintos buscadores y encontraron que ninguno cubría más de un tercio de la Web indexable conocida y que la combinación de varios buscadores podía cubrir más del triple de páginas que un único sistema.
- Concluyen que el uso de metabuscadores era una solución para localizar información en la Web puesto que garantizaba la cobertura del mayor número posible de páginas.
  - Uno de los primeros metabuscadores, de hecho anterior a su estudio, fue *MetaCrawler*.
  - Los dos principales motivos para crear un metabuscador eran:
    1. la porción de la Web sobre la que trabaja cada buscador es distinta del resto obligando a los usuarios a repetir la consulta en distintos buscadores,
    2. gran parte de los resultados son irrelevantes o enlaces "muertos".
  - *MetaCrawler* pretendía dar solución al primer problema ofreciendo una interfaz única para los distintos buscadores (esto es, distintas bases de datos) y el segundo filtrando los resultados recibidos.
- Aunque los metabuscadores tuvieron su momento lo cierto es que no eran imprescindibles y tanto el problema de la cobertura como el de la "frescura" de los enlaces eran resolubles por parte de los buscadores.

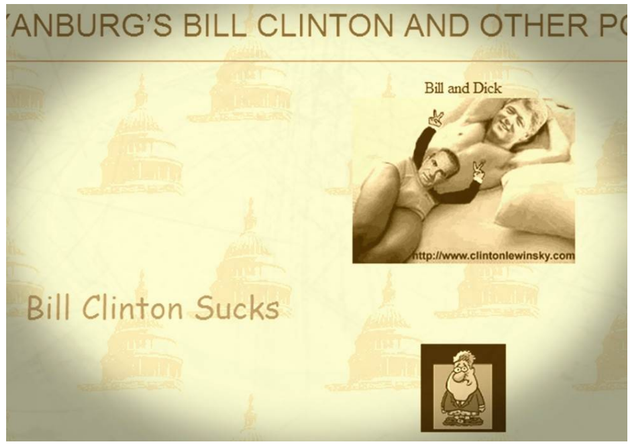
Estado de la Web pre-1998

> *I like the Internet. Really, I do. Any time I need a piece of crap shareware or I want to find out the weather in Bogota, I'm the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd. Somehow, whenever I'm looking for something in particular, I get 404s right across the board.*

The Straw Men (Michael Marshall, 2002)

Estado de la Web pre-1998

- En noviembre de 1997 sólo uno de los cuatro mayores buscadores comerciales podía encontrarse a sí mismo (retornar su propia página entre los primeros 10 resultados en respuesta a una consulta consistente en su nombre).
- En esa misma época un buscador podía retornar para la consulta `Bill Clinton` lo siguiente...



Estado de la Web pre-1998

- ¿Por qué si había cuatro décadas de investigación en recuperación de información funcionaba todo tan mal?
- Por el tamaño de la Web. En 1998 se había estimado que una cota inferior era de 320 millones de documentos. En cambio, la colección más grande de la época, la VLC de TREC tenía 7,49 millones de documentos y ocupaba solo 20.14 GB.
- Por la heterogeneidad de los documentos y de los usuarios de los buscadores. Mientras que en la investigación en recuperación de información se trabajaba con colecciones homogéneas y usuarios profesionales, cualquiera podía publicar en la Web y buscar información en la misma. Además, las consultas que construían los usuarios eran muy cortas (tres términos o menos).
- Porque, a diferencia de un sistema IR en una organización, la Web era un entorno adversarial y existían múltiples actores maliciosos que pretendían alterar el funcionamiento de los buscadores para posicionar sus sitios web. En aquella época el spam de palabras clave era particularmente habitual y dañino.
- Y, además, **porque la Web no era una colección gigante de textos sino que era un hipertexto gigante...**

La Web es un hipertexto gigante

- La Web era un sistema hipermedia distribuido pero todo el mundo pareció olvidarlo durante ¡9 años! e insistieron en tratarla como una colección de textos.
- En 1997 Massimo Marchiori hace una propuesta "alocada", ¡olvidemos el texto!

> *A great problem with **search engines' scoring mechanisms** is that they tend to score text more than hypertext.*
>
> *[...] **focusing separately on the "textual" and "hyper" components.***
>
> *The presence of links in a Web object clearly augments the informative content with the information contained in the pointed Web objects.*
>
> ***Recursively, links present in the pointed Web objects further contribute, and so on.** Thus, in principle, the analysis of the informative content of a Web object A should involve all the Web objects that are reachable from it [...]*
>
> ***This is clearly unfeasible in practice**, so, for practical reasons, we have to stop the analysis at a certain depth [...]*

Ranking basado en hiperenlaces

- Marchiori proponía una solución recursiva para determinar la relevancia de un sitio/página web en función de la topología del grafo.
- Todo algoritmo recursivo admite una versión iterativa.
- En 1998 se propusieron **¡dos!** soluciones iterativas diferentes:
  - El algoritmo *HITS* de Jon Kleinberg.
  - El algoritmo *PageRank* de Brin y Page.
- En ambos casos se calcula una medida de centralidad en la Web que tan solo tiene en cuenta los enlaces y que ignora completamente el texto de las páginas. Es obvio que, de ser efectivos, estos algoritmos serían inmunes a las técnicas de *webspam* existentes en aquella época.

Ranking basado en hiperenlaces

HITS (Hyperlink-Induced Topic Search)

- En 1998 Jon Kleinberg sentó las bases en las que sobre las que se apoyan los modernos buscadores al plantearse la viabilidad de un método algorítmico para estimar la relevancia de un documento, algo que según él era una característica subjetiva.
- Para ello definió los conceptos de **"autoridad" y hub** (concentrador). Una autoridad es un documento al que enlazan muchos otros puesto que, según Kleinberg, cada enlace recibido es un **"voto"** emitido por el individuo que estableció dicho enlace. Analizando el texto empleado en los enlaces puede determinarse el contexto en el cual el documento enlazado es una autoridad.
- Por su parte, un concentrador será un documento que contiene enlaces a muchas autoridades y es, por tanto, un recurso valioso para localizar información relevante en la Web.

Ranking basado en hiperenlaces

HITS (Hyperlink-Induced Topic Search)

- Estos conceptos fueron probados por Soumen Chakrabarti (1998a) y (1998b) mediante varios prototipos que tenían como objetivo localizar únicamente los documentos más relevantes para cada consulta, esto es, las autoridades.
- Para evaluar el rendimiento de estas técnicas se realizaron una serie de consultas genéricas empleando dichos prototipos, Yahoo! (un directorio) y Altavista (un buscador basado en robots) obteniendo, en cada caso, los diez documentos más relevantes.
- Posteriormente, un grupo de usuarios evaluó de manera "ciega" cada documento y valoró su relevancia en relación con la consulta planteada.
- **La relevancia media de los resultados proporcionados empleando la técnica de Kleinberg superaba el 50% frente al 40% de Yahoo! y el 20% de Altavista** abriendo la posibilidad de construir automáticamente taxonomías de documentos similares a las construidas por expertos humanos.
- **Curiosamente, la técnica no se desarrolló comercialmente...**

Ranking basado en hiperenlaces

PageRank

- En 1998 comenzó a operar el buscador, tal vez, más popular de la actualidad: Google.
- Éste, al igual que los motores de búsqueda "tradicionales", empleaba robots para explorar la Web en búsqueda de documentos pero, al contrario que estos, utiliza una técnica mucho más sofisticada para organizar los resultados de las consultas de los usuarios: el algoritmo *PageRank*, similar en ciertos aspectos al propuesto por Kleinberg.
- Al igual que Kleinberg, *PageRank* se basa en el uso de autoridades; sin embargo, no todos los enlaces son valorados del mismo modo sino en función de un valor numérico otorgado a cada documento, denominado también *PageRank*.
- Dicho valor indica el "prestigio" o la relevancia del documento y se propaga de unos documentos a otros: el *PageRank* de una página se divide por el número de enlaces de salida y se "transfiere" a los documentos enlazados.
- Así, documentos que reciben muchos enlaces aunque de poco valor serán muy relevantes y documentos que reciben pocos enlaces pero desde páginas con *PageRank* elevado serán igualmente importantes.

Ranking basado en hiperenlaces

PageRank

Ranking basado en hiperenlaces

Algunas características interesantes de *PageRank*

- Los valores de *PageRank* calculados para los nodos se **"estabilizan" con rapidez** (p.ej. 52 iteraciones son suficientes para obtener valores razonables para 322 millones de enlaces)
- Es **relativamente insensible a los valores de "partida"**, afectaría al número de iteraciones necesarias y a los valores finales (obviamente) pero no al ranking obtenido
- El *PageRank* total en la Web es constante
- Si el valor inicial asignado a cada documento es 1/N (número de documentos) el valor de *PageRank* equivale a la probabilidad de que un usuario llegue a dicho documento siguiendo enlaces al azar *(random surfer model)*.

Ranking basado en hiperenlaces

🤔 ¿Es verdad esto que pone la Wikipedia?

- En su artículo sobre buscadores se afirma lo siguiente:

> *In 1996, Robin Li developed the RankDex site-scoring algorithm for search engines results page ranking and received a US patent for the technology. It was the first search engine that used hyperlinks to measure the quality of websites it was indexing, predating the very similar algorithm patent filed by Google two years later in 1998.*

- La patente en cuestión es la número 5920859 y la pregunta es sencilla: ¿Es cierto que hubo un algoritmo anterior similar a PageRank (y HITS)?
- Leer esa patente y tratar de dirimir si, efectivamente, el algoritmo que describe es (o no) un algoritmo de centralidad en grafos. Si alguien se anima a hacerlo como ejercicio, por favor, que me escriba para que le asigne un "minipunto".

Problemas del ranking basado en hiperenlaces

> *[…] PageRanks are* ***virtually immune to manipulation*** *by commercial interests. For a page to get a high PageRank,* ***it must convince an important page, or a lot of non-important pages to link to it****. At worst, you can have manipulation in the form of buying advertisements (links) on important sites. But, this seems well under control since it costs money.*

Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, *The PageRank Citation Ranking: Bringing Order to the Web*

Lamentablemente, eso es mentira...

Problemas del ranking basado en hiperenlaces

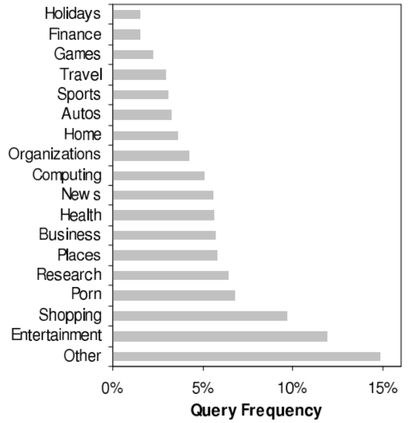
- Ya en 1998 se señalaron tres escenarios problemáticos:
  1. **Enlaces "nepotistas"** Cada enlace es un "voto" aunque provengan todos del mismo servidor.
     Davison, B.D. 2000, *"Recognizing Nepotistic Links on the Web"*
  2. **Enlaces automáticos** Todos estos algoritmos parten del supuesto que los enlaces son establecidos por un ser humano y eso no siempre es cierto *(Wordpress Website's Search Engine Spam)*
  3. **Documentos irrelevantes enlazados desde autoridades** Inevitable puesto que no hay ningún análisis de contenidos, sólo se emplea la topología del grafo
- Granjas de enlaces.
- *Google bombing*. No se tomaron medidas para desactivarlas hasta 2007 [1][2].
- **La Web seguía siendo un entorno adversarial**

¿Qué tal funcionaban los buscadores basados en ranking de hiperenlaces?

- Salvo en contadas ocasiones los buscadores comerciales no liberan sus *logs* de consultas por lo que la información sobre las consultas que hacen los usuarios y los resultados que reciben es fragmentaria.
- Prácticamente toda la información disponible fue obtenida por Amanda Spink y Jim Jansen trabajando con datos de *Excite*, *AlltheWeb* y *AltaVista* entre 1997 y 2002.
- Entre los principales resultados podemos destacar:
  - Los buscadores parecen adecuados para la mayor parte de usuarios. Según Jansen et al. (1998) y Silverstein et al. (1998) los usuarios de motores de búsqueda resuelven una necesidad de información con menos de dos consultas (en un 67% de los casos de acuerdo con el primer estudio y en un 78% según el segundo),
  - no suelen pasar de la primera página de resultados (en un 58% según los primeros y en un 85% según los segundos) y,
  - de acuerdo con Jansen y Spink (2003), un 66% de los usuarios examinan entre los resultados menos de 5 documentos y un 30% un único documento.
- Jansen y Spink argumentan que esto se debe a tres razones:
  1. las necesidades de información de la mayoría de internautas no son complejas,
  2. los primeros documentos retornados son realmente "autoridades" para la consulta formulada y,
  3. en promedio, alrededor del 50% de los documentos retornados son relevantes para una consulta específica desde la perspectiva del usuario.

¿Qué tal funcionaban los buscadores basados en ranking de hiperenlaces?

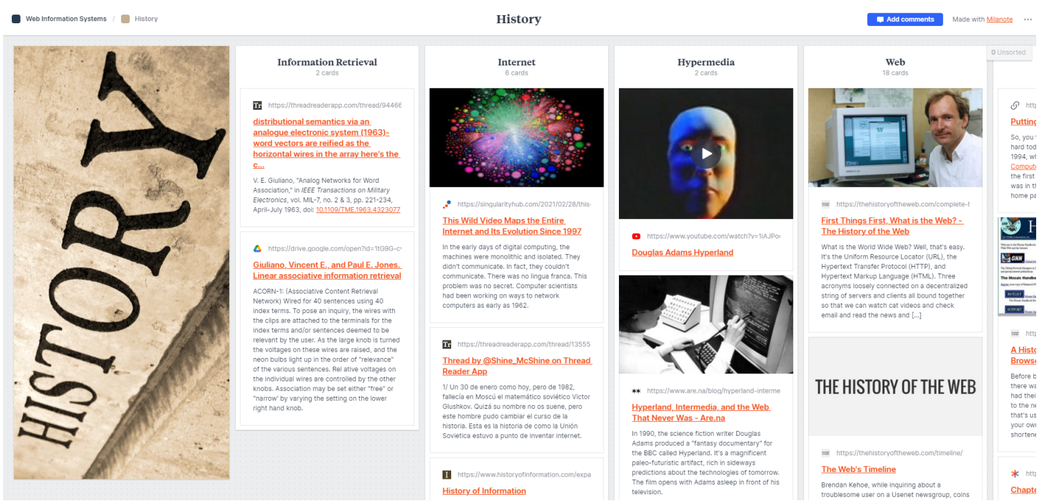
- Un estudio más reciente (2006) se llevó a cabo sobre un log de consultas del buscador de AOL que usaba el motor de Google.
- El espacio de consultas es inmenso (cientos de millones de consultas diferentes), cambiante y muy diverso. Sin embargo, menos de 10.000 consultas suponen ya el 25% de todo el volumen de consultas visto por un buscador y las consultas siguen siendo cortas (3.5 términos de media).
- El 85% de las consultas pueden clasificarse en tan solo 17 categorías.
- Las consultas navegacionales suponen el 21% del volumen total de consultas.
- El 41% de los usuarios usan el buscador una única vez al día.
- Existe una fracción muy pequeña de *"heavy users"* que hacen la mayoría de consultas y tienen una percepción de la calidad de los resultados muy distinta a la de la mayoría de usuarios.



¿Qué tal funcionaban los buscadores basados en ranking de hiperenlaces?

- Aparentemente, los buscadores funcionaban bien para la inmensa mayoría de usuarios la mayoría del tiempo.
- En realidad, los buscadores basados en ranking de hiperenlaces resuelven muy bien aquellas consultas para las que existen una o más páginas “autorizadas”.
- No obtienen tan buenos resultados cuando no existen tales autoridades. En este último caso, el usuario simplemente recibe una avalancha de información. (En la primera práctica se trabaja con distintos tipos de consultas para ver el funcionamiento de un buscador).

Para saber más...



Distintas propuestas para luchar contra la sobrecarga de información

- La cantidad de información disponible en Internet es enorme y los buscadores no son capaces de resolver todas las consultas satisfactoriamente.
- Algunas consultas obtienen como resultado cientos o miles de documentos sin existir ninguno adecuado entre los primeros aunque es razonable suponer que alguno de los restantes sí es relevante para las necesidades del usuario.
- El problema radica en encontrar en un conjunto de documentos muy grande unos pocos que sean de interés para el usuario.
- A lo largo de los 1990s se realizaron toda una serie de investigaciones sobre este asunto no sólo en la Web sino también en otros servicios como correo electrónico o grupos de *USENET*.
- Estos trabajos emplearon, de forma independiente o combinada, tres técnicas básicas:
    1. **agentes software**,
    2. **filtrado colaborativo** y
    3. **recomendación por contenidos**.

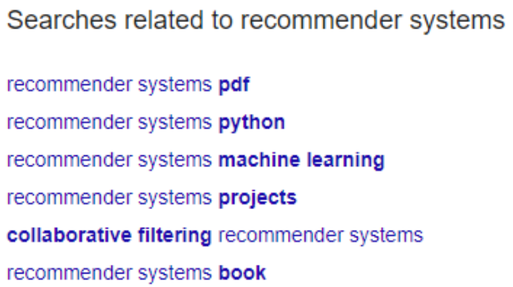
Distintas propuestas para luchar contra la sobrecarga de información

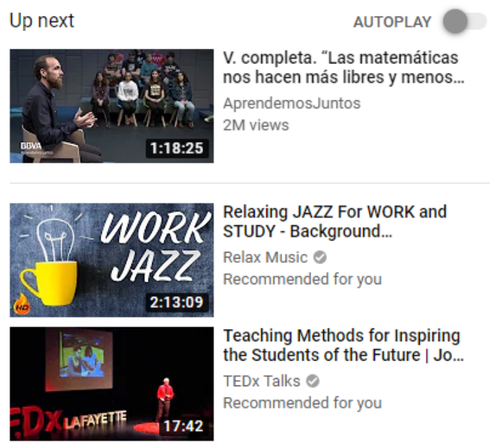
- Un **agente** es un elemento software capaz de interactuar con su entorno (incluidos otros agentes) para realizar una tarea en representación de un usuario o de otro agente.
  - Los agentes implementan algún tipo de inteligencia artificial que les permite actuar de manera autónoma y determinar las acciones apropiadas para responder a los eventos del entorno.
  - Los agentes vuelven a estar de moda bajo la apariencia de asistentes virtuales. Los más conocidos son Siri, Cortana, Alexa y Google Assistant.
  - Los modernos agentes utilizan técnicas de reconocimiento y generación de voz, procesamiento de lenguaje natural y respuesta a preguntas.
  - Muy recientemente Facebook anunció *Blender Bot 2.0*, un chatbot que podría resolver cuestiones de los usuarios consultando la Web. Para saber más: Principales motores para chatbots.
- El **filtrado colaborativo** proporciona a un usuario lo que otros individuos similares encontraron de utilidad antes que él. Un ejemplo típico es el servicio de *Amazon **"Customers who bought this item also bought…"***("Los clientes que compraron este producto también compraron").
- Por su parte, la **recomendación por contenidos** tiene como objetivo proporcionar documentos similares a un documento de partida y precisa, por tanto, de algún tipo de análisis del texto de los documentos.

Distintas propuestas para luchar contra la sobrecarga de información

Sistemas de recomendación

- Los sistemas de recomendación (por filtrado colaborativo, basados en contenidos o híbridos) son prácticamente ubicuos a día de hoy. En buscadores, en redes sociales, en portales multimedia, ...





Distintas propuestas para luchar contra la sobrecarga de información

Sistemas de recomendación

- Por diversas razones (fundamentalmente sesgos en los usuarios y maximización del *engagement* en las plataformas) los sistemas de recomendación **pueden tener efectos perniciosos**.
- El menor de los problemas son las denominadas *filter bubbles* (burbujas informativas) en las que los usuarios podrían acabar encerrados inconscientemente en base a los medios que leen y los amigos que sigen en redes sociales. [1][2]
- Mucho más dañinas son la difusión de desinformación o la radicalización de personas que pueden llegar a cometer actos violentos.
- Ejemplo de recomendación peligrosa en buscadores: `vaccines` > `vaccines pros and cons` > `why vaccines are bad | disadvantages of vaccines | vaccines to avoid`
- Hilo muy interesante sobre radicalización a través de YouTube.
- Theirtube, un simulador de *filter bubbles* para distintos tipos de personas: negacionistas del cambio climático, conspiranoicos, survivalistas, etc.
- Para saber más: aquí y en el *moodboard*.

La Web Semántica

- Paralelamente al desarrollo de técnicas como las de Kleinberg o Google para localizar documentos en la Web y al mismo tiempo en que se buscaban soluciones al problema de la sobrecarga de información en Internet, Tim Berners-Lee esbozaba el concepto de Web Semántica que, junto con James Hendler y Ora Lassila, refinaría posteriormente.

- **El objetivo básico de la Web Semántica era permitir que agentes software fueran capaces de "consumir" documentos disponibles en la Web para inferir nuevo conocimiento**. Para ello los documentos deberían construirse empleando **lenguajes "semánticos"** que permitirían no sólo **anotar metainformación sino también especificar las relaciones existentes entre los metadatos**. Para realizar esa labor se optó por la utilización de **ontologías** (no exentas de polémica [1][2]).

La Web Semántica

> *Una ontología es la especificación de una conceptualización. Esto es, una descripción de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes*

Gruber (1993)

> *Un documento o fichero que define formalmente las relaciones entre términos. Una ontología típica para la Web consta de una taxonomía y de un conjunto de reglas de inferencia.*

Berners-Lee, Hendler y Lassila (2001)

La Web Semántica

pre-Web-Semántica

- Con anterioridad o simultáneamente a la propuesta de Berners-Lee para la Web Semántica ya se estaban realizando una serie de trabajos que tenían como objetivo desarrollar lenguajes que permitiesen definir tales ontologías y utilizarlas para etiquetar documentos en la Web, lo que podríamos denominar **"pre-Web-Semántica"**.
- Cabe destacar los proyectos *SHOE*, *WebKB* y *Ontobroker*/*On2broker*.

La Web Semántica

pre-Web-Semántica

- *SHOE* fue una de las primeras iniciativas destinadas a proporcionar un lenguaje de marcado semántico.
- Se trata de una extensión del lenguaje HTML que permite desarrollar ontologías y utilizar las clases y relaciones definidas en una o más de esas ontologías para marcar zonas específicas de un documento HTML.
- Luke *et al.* describen asimismo una herramienta, *Exposé*, que explora la Web en busca de páginas anotadas con *SHOE* y almacena los asertos que encuentra en una base de conocimiento que puede utilizarse posteriormente para realizar consultas.

La Web Semántica

pre-Web-Semántica

- *WebKB* tenía como objetivo construir, de forma automática, una base de conocimiento que reflejase el contenido de la Web de una forma inteligible para una máquina.
- Para lograr esto el sistema debía recibir una ontología que describiese las clases y relaciones, así como un conjunto de documentos, etiquetados sobre la base de dicha ontología, que servirían como conjunto de entrenamiento.
- Así, tras un período de entrenamiento adecuado, el sistema sería capaz de procesar documentos HTML y producir documentos marcados semánticamente de acuerdo a la ontología de partida.
- En ciertos aspectos, la tarea que resolvía *WebKB* era similar a las que tratan de resolver los sistemas de reconocimiento de entidades y de extracción de información. *OpenCalais* ofrece un servicio web de extracción de entidades pudiendo ofrecer la salida en RDF

La Web Semántica

pre-Web-Semántica

- *Ontobroker* fue una iniciativa muy similar a *SHOE* puesto que proponía una serie de herramientas para definir ontologías, etiquetar documentos basándose en dichas ontologías y realizar consultas e inferencia sobre una base de conocimiento.
- Posteriormente evolucionaría hacia *On2broker* cuya principal novedad fue la utilización de tecnologías como *XML* o *RDF*.

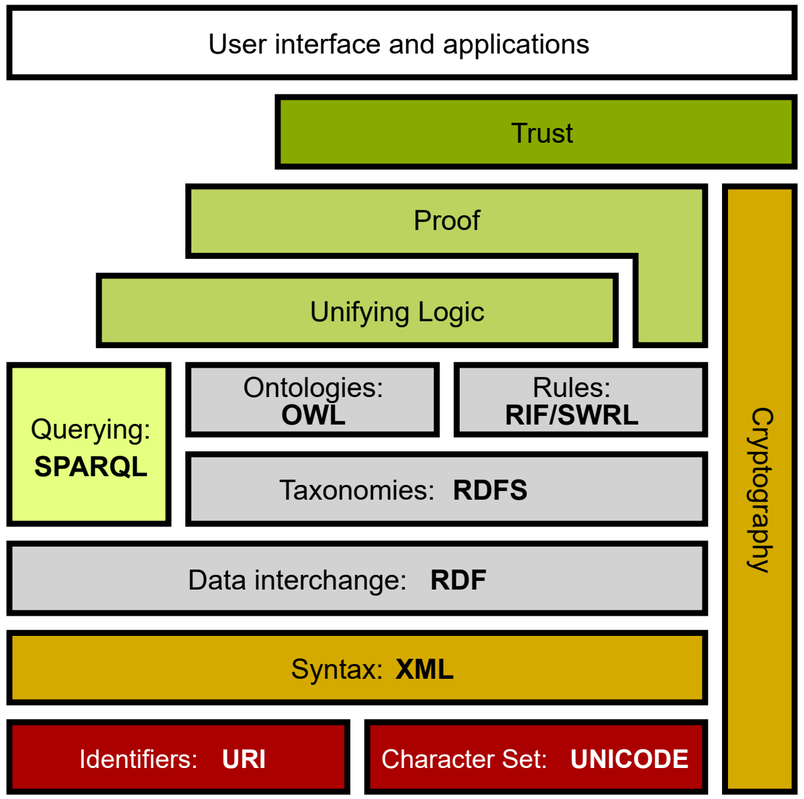
La Web Semántica

Pila de tecnologías

- XML y RDF constituyeron las bases iniciales sobre las que comenzar a construir la Web Semántica puesto que el primero posibilita la construcción de nuevos lenguajes de etiquetas, por ejemplo RDF que, a su vez, permite expresar asertos.
- Sin embargo, son necesarias toda una serie de capas encima de RDF para desarrollar finalmente la Web Semántica.
- Por ejemplo, aunque RDF permite dar valores a las distintas propiedades de diferentes recursos no dispone de mecanismos para describir esas propiedades ni para describir las relaciones entre las propiedades y otros recursos.
- Para ello es necesario un lenguaje que permita definir vocabularios RDF. Dicho lenguaje, construido mediante RDF, es RDF Schema o RDF(S). Este lenguaje define clases y propiedades que permiten, a su vez, describir nuevas clases, propiedades y recursos.
- Sin embargo, tampoco RDF ni RDF Schema son capaces por sí solos de modelar ontologías, razón por lo que comienzan a desarrollarse lenguajes para este fin análogos a los definidos durante la fase pre-Web-Semántica con la diferencia de que los nuevos lenguajes se construyen sobre el estándar RDF(S).
- Ejemplos de estas extensiones ontológicas para RDF Schema son la de Staab *et al.*, OIL o DAML-ONT.
- Posteriormente DAML-ONT y OIL convergieron en el lenguaje DAML+OIL que terminaría evolucionando hacia OWL, una recomendación del W3C y, por tanto, el estándar para la construcción de ontologías.

La Web Semántica

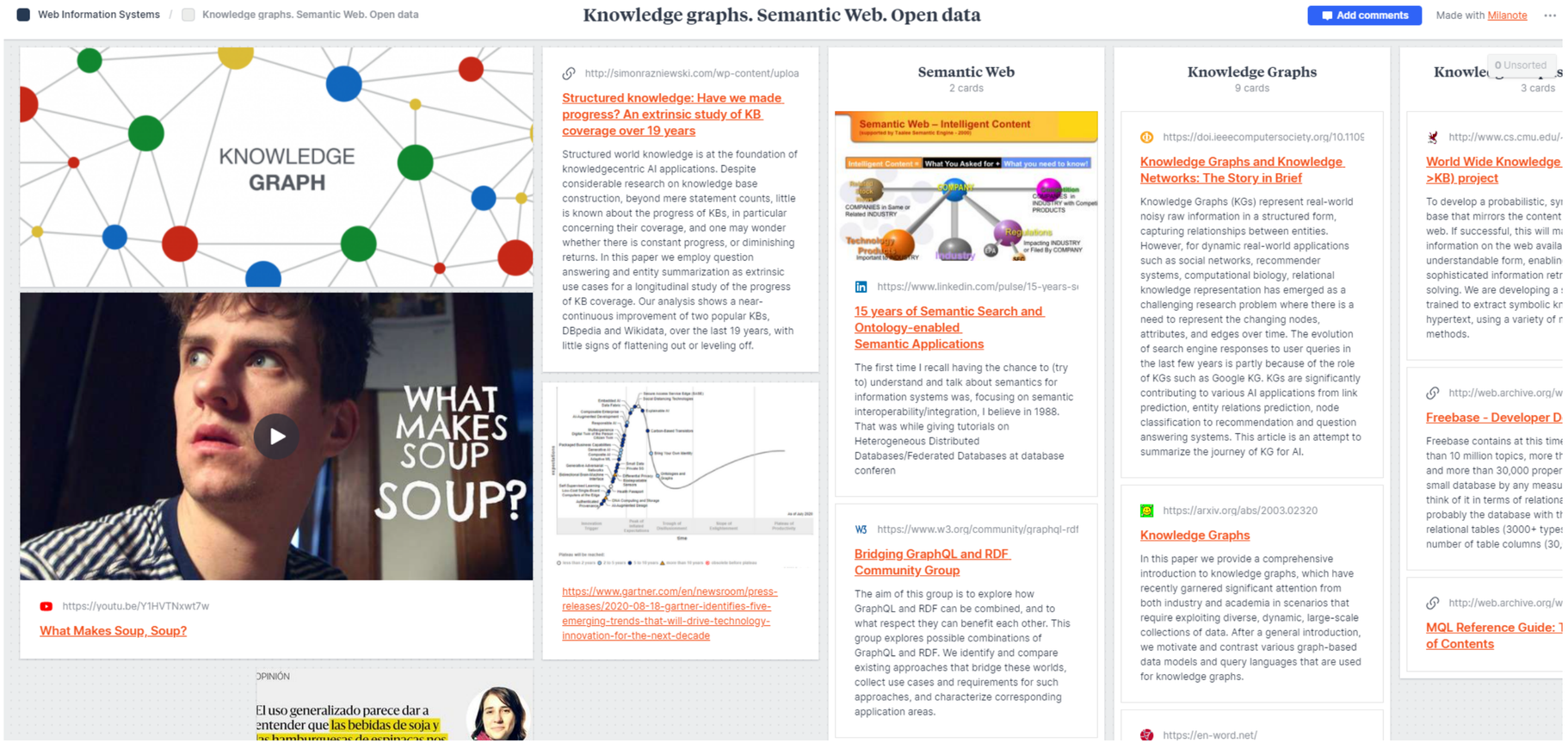
Pila de tecnologías



Consultas en la Web Semántica

- Al igual que había sucedido con la propuesta de la Web, la pila inicial de la Web Semántica carecía de un sistema de consulta aunque muy pronto se comenzaron a explorar varios:
  - *Metalog*,
  - *SquishQL/RDQL* o
  - *RQL/SeRQL*.
- En 2004 se constituyó el *RDF Data Access Working Group* en el que participaron, entre otros, autores de *Metalog*, *SquishQL*, *RDQL*, *SHOE* y *DAML-ONT*.
- Muy pronto se esbozaron los primeros casos de uso y un primer borrador del lenguaje de consulta *BRQL*.
- Dichos trabajos llevaron, en 2008, a la publicación de *SPARQL* como una recomendación del W3C. El estándar actual es *SPARQL 1.1*, de 2013.
- *SPARQL* permite seleccionar información, extraer subgrafos RDF y construir nuevos grafos RDF a partir del resultado de una consulta.
- Se trata de un lenguaje expresivo y potente, aunque su objetivo es resolver consultas fundamentalmente "metasemánticas", por ejemplo:
  - Encontrar la dirección de correo de Jonhny Lee Outlaw.
  - Encontrar en la web de un proveedor información sobre el repuesto de una pieza así como la lista de piezas que deben ser sustituidas junto con la defectuosa.
  - Recibir puntualmente información sobre libros, películas y música que cumplan unos criterios de título, precio y autor.
  - Grabar todos los programas de televisión sobre el jugador de béisbol Ichiro.

Para saber más...



Consultas en la Web Semántica

- La utilidad de *SPARQL* está fuera de toda duda.
- Sin embargo, el problema es resolver en la Web de manera adecuada y automática consultas informativas mucho más abiertas y ambiguas, formuladas en cualquier lenguaje natural (tal vez con errores tipográficos, ortográficos o gramaticales) y susceptibles de sobrecargar de información al usuario. Por ejemplo:
  - `history and cultural Bengal`
    (historia y Bengal cultural).
  - `accurate predictors of aspiration pneumonia: how important is dysphagia?`
    (predictores adecuados para la neumonía por aspiración: ¿cuán importante es la disfagia?)
  - `degenerative disk disease`
    (enfermedad degenerativa de disco). En el contexto médico es más común la forma "disc" que "disk".
  - `muscel` *(sic)* `aches during pregnancy`
    (dolores muscualres (sic) durante el embarazo). Consulta con error tipográfico.

Google

history and cultural Bengal

All Maps Images News Videos More Settings Tools

About 24,600,000 results (0.94 seconds)

Bengal has a recorded **history** of 1,400 years. The **Bengali** people are its dominant ethnolinguistic tribe. ... **Bengal** was the richest part of Medieval India and hosted the subcontinent's most advanced political and **cultural** centers during the British Raj. The partition of **Bengal** left its own **cultural** legacy.

Culture of Bengal - Wikipedia
https://en.wikipedia.org/wiki/Culture_of_Bengal

About this result Feedback

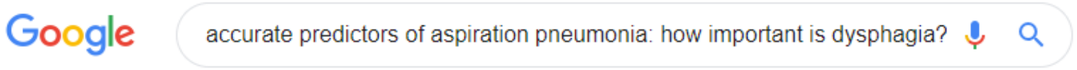
People also ask

- What is the Bengali culture?
- Who was the founder of Bengal?
- Where did Bengali originate?
- Who was the first king of Bengal?

Feedback

Searches related to history and cultural Bengal

- **bengali culture** and **lifestyle**
- **bengali culture facts**
- **culture** and **tradition of west** bengal **ppt**
- **essay on west** bengal
- **west** bengal **culture images**
- history **of west** bengal
- **bengali culture marriage**
- **traditional dance of west** bengal



All Images News Videos Shopping More Settings Tools

About 86 results (0.53 seconds)

Aspiration pneumonia is a major cause of morbidity and mortality among the elderly who are hospitalized or in nursing homes. ... **Dysphagia** was concluded to be an **important** risk for **aspiration pneumonia**, but generally not sufficient to cause **pneumonia** unless other risk factors are present as well.

Predictors of aspiration pneumonia: how important is dysphagia?
https://www.ncbi.nlm.nih.gov/pubmed/9513300

About this result Feedback

Predictors of aspiration pneumonia: how important is dysphagia?
https://www.ncbi.nlm.nih.gov/pubmed/9513300 ▾
by SE Langmore - 1998 - Cited by 853 - Related articles
Logistic regression analyses then identified the **significant predictors of aspiration pneumonia**. ... **Dysphagia** was concluded to be an **important** risk for **aspiration pneumonia**, but generally not sufficient to cause **pneumonia** unless other risk factors are present as well.
Missing: ~~accurate~~ | Must include: accurate

Searches related to accurate predictors of aspiration pneumonia: how important is dysphagia?

predictors of aspiration **pneumonia:** how important is **dysphagia? pdf**
predictors of aspiration **langmore**
aspiration pneumonia **in nursing homes**
aspiration pneumonia **speech therapy**
langmore 2002 study
prandial aspiration

Google

degenerative disk disease

All Images Videos News Shopping More Settings Tools

About 1,450,000 results (0.68 seconds)

Degenerative disk disease is when normal changes that take place in the **disks** of your spine cause pain. Spinal **disks** are like shock absorbers between the vertebrae, or bones, of your spine. They help your back stay flexible, so you can bend and twist. As you get older, they can show signs of wear and tear. Dec 17, 2017

Degenerative Disk Disease: Symptoms, Causes, Diagnosis, Treatment
https://www.webmd.com/back-pain/degenerative-disk-disease-overview

About this result Feedback

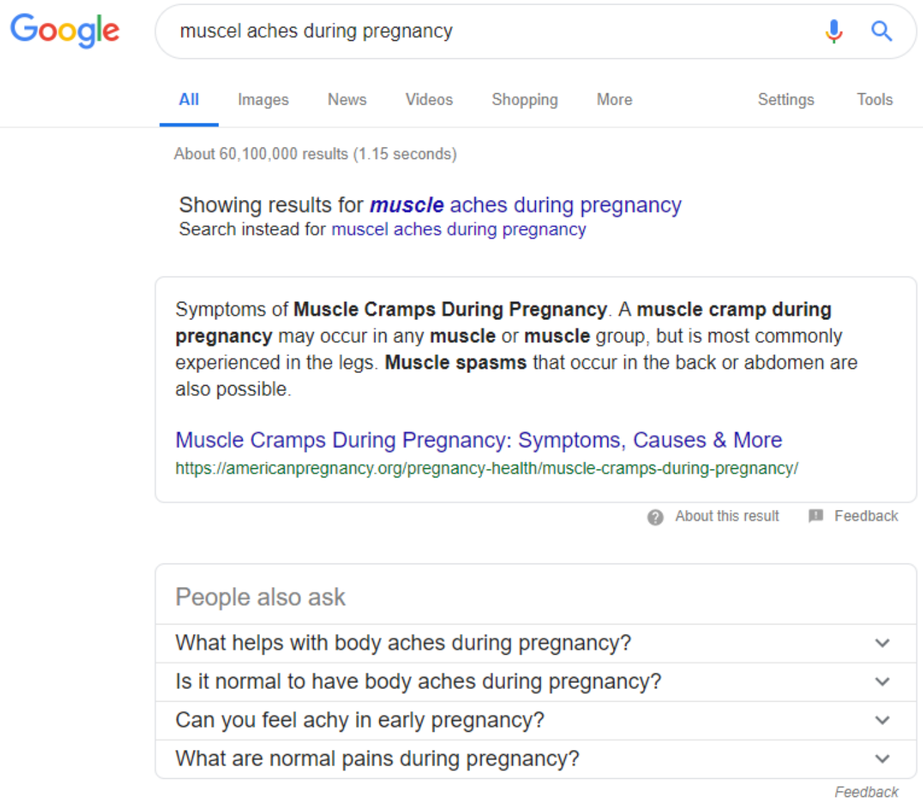
People also ask

- What is the best treatment for degenerative disc disease?
- Is degenerative disc disease considered a disability?
- What causes degenerative disk disease?
- Can degenerative discs heal?

Feedback

Searches related to degenerative disk disease

- degenerative **disc** disease **treatment**
- degenerative **disc** disease **symptoms**
- degenerative **disc** disease **causes**
- degenerative **disc** disease **pictures**
- degenerative **disc** disease **in neck**
- degenerative **disc** disease **exercises**
- degenerative **disc** disease **l5-s1**
- degenerative **disc** disease **surgery**



Searches related to muscle aches during pregnancy

arm muscle **pain** during pregnancy

leg muscle **twitching** pregnancy

stomach muscle **spasms** during pregnancy

leg cramps during pregnancy **while sleeping**

home remedies for leg cramps during pregnancy

muscle **twitching** during pregnancy

whole body pain during pregnancy

back muscle **spasms** pregnancy

Los buscadores no se han dormido...

- Han transcurrido casi dos décadas desde la propuesta inicial de la Web Semántica. En ese tiempo los buscadores han incorporado multitud de características nuevas más allá de los resultados orgánicos, como se ha podido ver en las consultas anteriores.
- En todas esas consultas la información que probablemente necesita el usuario **no** está necesariamente en los resultados sino en otros elementos ofrecidos por el buscador. Tales como:
  - ***Featured snippets***. Pequeño resumen obtenido de una de las páginas de resultados ofrecido como una tarjeta. Es similar a los *Knowledge panels* y *Knowledge cards* que extraen información del *Knowledge Graph*. (Parte de los datos de este grafo provenían de *Freebase* que tenía muchos puntos en común con las tecnologías de la Web Semántica y con *Linked Data*).
  - **Preguntas relacionadas**. Lista de preguntas generadas algorítmicamente con un pequeño resumen como respuesta.
  - **Consultas relacionadas**. No confundir con el anterior, se trata de consultas obtenidas por minería de los logs de consultas y que están relacionadas con la consulta de partida.
- Además, la corrección de errores orto/tipográficos es constante.

Para saber más [1][2] y en el *moodboard*

Minería Web

- Una parte importante de las características anteriores se han obtenido mediante **minería web**.
- La minería web consiste en la extracción de nuevo conocimiento a partir de datos disponibles en la Web como de datos generados con su uso diario.
- La minería web puede dividirse en tres grandes áreas:
  1. Extracción de conocimiento a partir de la **estructura hipertextual** de la Web (p.ej. algoritmos PageRank y HITS); es decir, explotar que la **Web es un grafo**.
  2. Extracción de conocimiento a partir del **uso de la Web** (p.ej. logs de servidores y buscadores, anécdota y libro no técnico).
  3. Extracción de conocimiento a partir de los **contenidos disponibles** en la Web (la Web como *corpus*)
- La minería web es **multidisciplinar**, involucra aprendizaje automático, procesamiento de lenguaje natural, estadística, recuperación de información, bases de datos o tecnologías de web semántica...
- Para saber más:
  - *Mining the Web* (Soumen Chakrabarti).
  - *Mining of Massive Datasets* (Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman)
  - *Web Data Mining* (Bing Liu)
  - *Mining the Social Web* (Matthew A. Russell)

La Web como *corpus*

- Un *corpus* es una **colección de documentos** que muestran el uso real de la **lengua natural**.
- Pueden ser **monolingües o multilingües** y estos, a su vez, **paralelos o comparables**.
- Los *corpora* multilingües son un recurso fundamental para la construcción de sistemas estadísticos de traducción automática.

OpenNMT is an open source (MIT) initiative for neural machine translation and neural sequence modeling.

- No es extraño utilizar recursos obtenidos de la Web para entrenar clasificadores de diversas temáticas. (Recordad, por ejemplo, los ejercicios de Repositorios de Información con Reddit pero también [1][2][3][4])

La Web en tiempo real

- La Web en tiempo real hace referencia a diversos sitios en los que la información se genera de manera continua y es consumida por los usuarios en tiempo cuasi-real.
- Sitios web como *Instagram*, *Facebook* o *Twitter* son la mejor representación de esta Web aunque podrían incluirse también blogs, sitios de noticias y foros.
- La minería web aplicada en la Web en tiempo real plantea multitud de retos:
  - Escala.
  - Tiempo de reacción.
  - Frescura de la información.
  - Relevancia.
  - Reputación/confiabilidad en los autores (generalmente usuarios).
  - Posibilidades de intoxicación informativa.
  - Pertinencia para el usuario que hace la consulta (personalización, geolocalización, etc.)

Linked [Open] Data

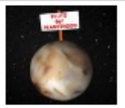
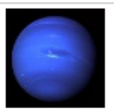
- En 2006 Tim Berners-Lee acuñó el término *Linked Data*.
- El objeto fundamental de *Linked Data* es la publicación (y enlace) de datos estructurados mediante RDF y su consulta semántica mediante SPARQL.
- La relación conceptual entre *Linked Data* y Web Semántica es "complicada"...
  - El propio Bernes-Lee definió *Linked Data* como *"Semantic web done right"*.
  - Hay quien lo considera un simple ejercicio de *rebranding*.
  - Según el W3C *Linked Data* es un conjunto de buenas prácticas para la publicación de datos estructurados en la Web; la Web Semántica, en cambio, sería la pila de tecnologías que permitirían la construcción de *Linked Data* además de su explotación.
  - Para saber más: [1][2] y el *moodboard*.
- Para complicar las cosas un poco más se introdujo el concepto de "Datos abiertos" que pueden estar enlazados o no. Los datos abiertos son una filosofía y práctica que persigue que ciertos datos estén disponibles de manera libre para su reutilización y remezcla por parte de cualquiera.

La Web como base de datos

- A simple vista podría parecer que recuperación de información en la Web, minería Web y Web Semántica/*Linked Data* persiguieran objetivos no enteramente alineados.
- En realidad todos persiguen (usando distintos medios) que se pueda explotar la Web como si se tratase de una base de datos gigantesca.
- Así, mientras que en algunas propuestas se enfatiza el marcado semántico de metadatos, en otras se trata de lograr la extracción de información estructurada a partir de información poco o nada estructurada.
- Para saber más:
  - *Information extraction and Google Squared*.
  - *Google Squared* (discontinuado).

Google squared labs planets Square it Add to this Square

planets

	Item Name	Image	Description	Orbital Period	Equatorial Surface	Mean Density	Add columns Add
	Earth		Web, www.nineplanets.org. Introduction and FAQ · What's New · Express Tour · Solar System Overview · The Sun · Mercury · Venus · **Earth** ...	365.256366 days	9.780327 m/s2	5.515	
	Jupiter		**Jupiter** is classified as a gas giant along with Saturn, Uranus and Neptune. Together, these four planets are sometimes referred to as the Jovian planets. ...	4331.572 days	No value found	11.86	
	Pluto		Originally classified as a planet, **Pluto** is now considered the largest member of a distinct population called the Kuiper belt. ...	90 613.305 days	No value found	1.95 gm/cm	
	Saturn		While you're shopping for **Saturn** vehicles, we've given you an easy way to keep information for the next time you visit **Saturn**.com. With My Saved Info, ...	29.46 yrs	8.96 m/s2	0.7*	
	Mercury		Planet **Mercury** is too small for its gravity to retain any significant atmosphere over long periods of time. The weak atmosphere contains hydrogen, helium, ...	88 days	3.7 m/s² 0.38 g[3]	5.427 g/cm³[3]	
	Ceres		Had this resolution been adopted, it would have made **Ceres** the fifth planet in order from the Sun. However, it was not accepted, and in its place an ...	4.60 yrs	0.27 m/s²	No value found	
	Neptune		**Neptune** is 17 times the mass of Earth and is slightly more massive than its near -twin Uranus, which is 15 Earth masses and not as dense. ...	164.79	11.0	1.64	
	Add items Add	or Add next 10 items					

Not finding the right items? Start with an empty Square.

Google squared labs

Rock Music Genres | Square it | Add to this Square

Sign in to save your Squares

Rock Music Genres

17 items

Item Name	Image	Description	Contact	Typical Instrumen	Mainstream Popul
Progressive rock		**Progressive rock** is often lumped together with other similar genres like art **rock**, symphonic **rock**, and **progressive** heavy metal. The artists try to take the ...	INSULT TO TRADITION	Guitar – Bass – Keyboards – Piano – Drums – optionally vocals, and other	High in the 1970s, revival in the 1980s, moderate in the 1990s, and a
Metal		Offers weekly podcasts dedicated to heavy **metal** music. ... Website dedicated to all **metal** genres. Includes interviews, editorials, message board, ...	John Terry	Electric guitar - Bass guitar - Drums - Vocals - Keyboards	Heavy
Garage punk		The **GaragePunk** Podcast Network's newest podcast host, the Barman (from the I-94 Bar), brings some all-Australian Real Rock Action to ya in an hour-long ...	No value found	Guitar - Bass - Drums - Keyboard	No value found
Rockabilly		Traces the history and highlights of the music. Includes performances. Features world's largest gathering of **rockabilly** entertainers and history makers. ...	Login / Logout	Guitar - Double bass - Drums - Piano	No value found
Boogie rock		**Boogie rock** is a music genre which came out of the hard heavy blues-**rock** of the late 1960s. It tends to feature a repetitive driving rhythm in place of ...	No value found	Guitar - Bass - Drums - Piano - Harmonica	Peaked in the 1970s in Europe and the Americas
Industrial		Throbbing Gristle opposed the elements of traditional rock music remaining in the punk rock scene, declaring **industrial** to be "anti-music. ...	James Loughran	No value found	Low
Blues rock		The first **rock** and roll was an offshoot of the **blues** genres by artists like Fats Domino and Elvis. The next generation in the sound were garage **rock** and ...	703.362.0549	Electric guitar, Bass guitar, drums, vocals, Hammond organ, Harmonica	Peak popularity in th 1960s and 1970s.
Psychedelic rock		**Psychedelic rock** is a style of **rock** music that is inspired or influenced by **psychedelic** culture, and attempts to replicate and enhance the mind-altering ...	No value found	Electric guitar (usually with guitar effects such as fuzz, phaser, flanger, reverb etc.) -	Peaked in the late 1960s
Southern rock		Not long after **southern rock** began to	Ph: 61 8 8682 1699	No value found	No value found

Google squared labs

united states national parks | Square it | Add to this Square

Sign in to save your Square

united states national parks — 7 item

Item Name	Image	Description	Nearest City	Established	Rooms	Add columns
Channel Islands National Park		Close to the California mainland, yet worlds apart, **Channel Islands National Park** encompasses five remarkable **islands** (Anacapa, Santa Cruz, Santa Rosa, **...**	Santa Barbara	March 5, 1980	1, 2, 3	
Yellowstone National Park		Established in 1872, **Yellowstone National Park** is America's first **national park**. Located in Wyoming, Montana, and Idaho, it is home to a large variety of **...**	West Yellowstone, Montana; Gardiner, Montana; Jackson, Wyoming	March 1, 1872	1 2	
Grand Canyon National Park		**Grand Canyon National Park**, Zoroaster and Brama Temples from the South Kaibab Trail **... Grand Canyon National Park** P.O. Box 129. **Grand Canyon**, AZ 86023 **...**	Fredonia, Arizona (North Rim) and Grand Canyon, Arizona	February 26, 1919	No value found	
Yosemite National Park		**Yosemite National Park**, one of the first wilderness parks in the United States, is best known for its waterfalls, but within its nearly 1200 square miles, **...**	Mariposa	October 1, 1890	No value found	
Golden Gate National Recreation Area		**Golden Gate National Recreation Area** is currently undergoing a multi-phased process to plan for accessibility improvements park-wide. **...**	San Francisco, California	October 27, 1972	No value found	
Hawaii Volcanoes National Park		Hawai'i **Volcanoes National Park** displays the results of 70 million years of volcanism, migration, and evolution -- processes that thrust a bare land from **...**	Hilo	August 1, 1916	No value found	
Death Valley National Park		**Death Valley National Park** is usually considered a winter **park**, but it is possible to visit here all year. When is the best time to visit? **...**	Pahrump, Nevada	February 11, 1933 (Monument) October 31, 1994	No value found	
Add items Add	or Add next 10 items					

Google squared labs

comic book character | Square it | Add to this Square

Saved Squares ▾ | Save

comic book character 7 items

	Item Name	Image	Description	Publisher	First Appearance	Author	Real Name	Add columns
☒	**Wolverine**		**Wolverine** is a comic book character that first appeared in The Incredible **.....** In 1936, **Wolverine** works alongside time-traveling Kitty Pryde and Rachel **...**	Marvel Comics	The Incredible Hulk #180	NineInchNail..	Wade Wilson	
☒	**Spider-Man**		**Spider-Man** is a fictional superhero in comic books published by Marvel Comics. **...** Marvel has featured **Spider-Man** in several comic book series, **...**	Activision	Amazing Fantasy #15	No value found	Peter Parker	
☒	**Batman**		In the modern era, the **Batman** "universe" puts out more monthly comic books than any other universe in comics, though Marvel's X-Men often rivals it in sheer **...**	DC Comics	Detective Comics #27	No value found	Bruce Wayne	
☒	**Scarlet Witch**		The **Scarlet Witch** (Wanda Maximoff) is a fictional character that appears in comic books published by Marvel Comics. The character first appears in X-Men #4 **...**	Marvel Comics	Uncanny X-Men #4	Loeb, Jeph.	Wanda Maximoff	
☒	**Superman**		**Superman**, the Man of Steel, is one of DC Comic's flagship characters. Reporter Clark Kent is sheepish, and works at the Daily Planet. As **Superman**, the last **...**	DC Comics,	Action Comics #1	No value found	Kal-El	
☒	**Venom**		**Venom**, or the **Venom** Symbiote, is a fictional character, a symbiote life **...** In 2009, **Venom** was ranked as IGN's 22nd Greatest Comic Book Villain of All Time. **...**	MARVEL COMICS	No value found	No value found	Edward Eddie Brock	
☒	**Punisher**		**Punisher**. **Punisher** is a comic book character that first appeared in The Amazing Spider-Man #129. Edit this Page Add to a list **...**	THQ	Amazing Spider-Man #129	Garth Ennis	Frank Castle	
	Add items Add	or Add next 10 items						

La Web como base de datos

Tecnologías relevantes

- Tecnologías semánticas (p.ej. RDF—bajo la forma de JSON-LD—y SPARQL).
- Extracción de información, de entidades, de **términos**.
- Respuesta de preguntas.
- **Recuperación de información**.
- **Resumen automático**.
- Traducción automática.

...

La Web tiene sesgos, es adversarial y efímera

- En ingeniería suelen cegarnos las posibilidades y tendemos a obviar los inconvenientes, problemas y efectos colaterales. *Move fast and break things.*
- Es fantástico usar la Web como un *corpus* gigantesco pero debemos ser conscientes de que tiene sesgos.
  - ¿Cuántas mujeres y miembros de minorías hay en esta colección? ¿Y en esta?
  - ¿Cuántas bodas no occidentales aparecen aquí?
  - ¿Cuántas parejas no convencionales aparecen en esta colección?
  - **Más ejemplos con consultas sobre historia de la humanidad.**

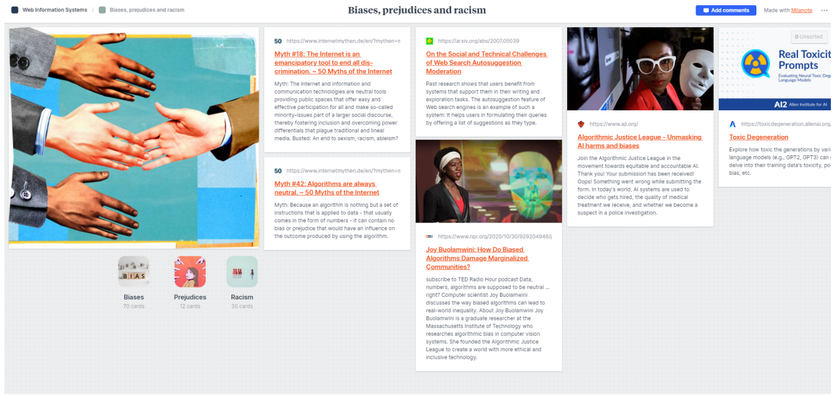
La Web tiene sesgos, es adversarial y efímera

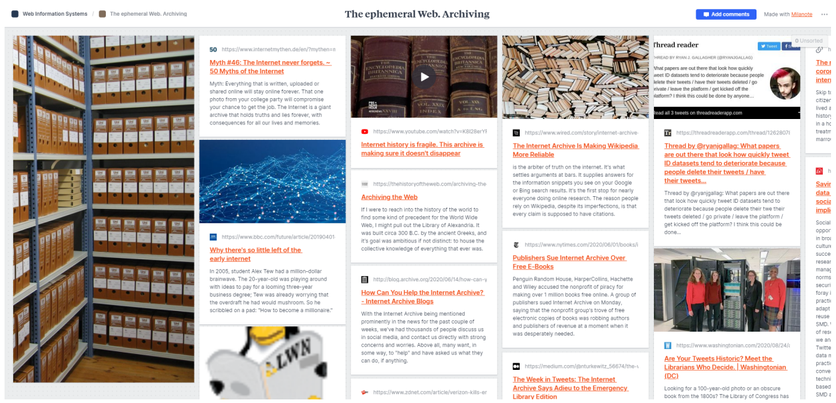
- La Web no solo sigue siendo un entorno adversarial (*spamdexing*), se ha convertido en un **campo de batalla propagandístico**. (Caso concreto de bots políticos en España)
- **No solo los humanos publican contenido en la Web.**
  - Desde hace algún tiempo existen los denominados ***Large Language Models***, sistemas de aprendizaje profundo entrenados sobre cantidades ingentes de texto procedente fundamentalmente de la Web y que pueden generar contenido nuevo que tiene una apariencia de verosimilitud.
  - Ese texto puede publicarse en la Web y ser utilizado por los buscadores para responder consultas y preguntas. Ejemplos: *How many bonks are in a quoit?* o *How many rainbows does it take to jump from Hawaii to seventeen?*
  - **Esos modelos tienden a adquirir los peores rasgos exhibidos por los textos de la Web respondiendo así de manera sesgada y prejuiciosa** (véase).

La Web tiene sesgos, es adversarial y efímera

- **La Web es mutable y no se ha diseñado para perdurar**. Cantidades ingentes de sitios web y contenidos han desaparecido sin dejar huella:
  - El 40% de los enlaces de la página del millón de dólares están rotos.
  - *Geocities*, que llegó a albergar 38 millones de páginas, fue cerrada por *Yahoo!* en 2009.
  - En octubre de 2014 cerraba *La Coctelera*, una plataforma de blogs que, en sus momentos de mayor expansión, llegó a tener cientos de miles de usuarios y alcanzar los 2 millones de posts.
  - En 2016 cerraba definitivamente *Tuenti*, la red social española con 20 millones de usuarios.
  - *MySpace* perdió 12 años de música subida por sus usuarios en una migración fallida.
  - En 2019 *Facebook* perdió accidentalmente publicaciones de Mark Zuckerberg, algunas de relevancia histórica. Afortunadamente, alguien había tenido la precaución de archivarlos.
  - *Yahoo!* cerró *Yahoo Groups* el 14 de diciembre de 2019 y *Yahoo Answers* el 4 de mayo de 2021 borrando todo el contenido.
  - Los foros de *Meristation* cerraron (y fueron borrados) el 5 de mayo de 2021.
  - Tarde o temprano *Facebook*, *Twitter* o *Instagram* desaparecerán sin dejar rastro...
  - **La *Wayback Machine* del *Internet Archive* trata de aliviar este problema al preservar capturas de sitios web a lo largo del tiempo**: en estos momentos almacena 585.000 millones de páginas web (visitar).

Para saber más...





Aquí y ahora...

- En 70 años la capacidad de almacenamiento ha aumentado de manera vertiginosa.
- La cantidad de datos producidos y almacenados por la humanidad es ingente.
- Las posibilidades de extraer conocimiento valioso son muchas.
- Las tecnologías disponibles son muy variadas, algunas con un rendimiento muy bueno.
- El *memex* de Vannevar Bush ya está disponible desde hace mucho tiempo.
- Lo que aún no tenemos es a *J.A.R.V.I.S.* de *Iron Man*.
- Por otro lado, no todo el mundo es bueno, la Web es un entorno adversarial plagado de sesgos y desinformación.
- A pesar de nuestra capacidad tecnológica, la Web es efímera, la mayor parte del contenido producido desaparecerá algún día sin dejar huella.
- **Es un área de trabajo apasionante. ¡Disfrutad de la asignatura!**