Resumen Tema 7: Dinámica de la web

José Alberto Benítez Andrades

Abril 2011

En este trabajo se resumen las conclusiones obtenidas después de haber realizado la lectura de los artículos propuestos M. Levene and A. Poulovassilis *"Web Dynamics"* , Ricardo Baeza-Yates, Bárbara J. Poblete y Felipe Saint-Jean. *Evolución de la web chilena,* Edward T. O'Neill, Brian F. Lavoie, Rick Bennett *"Trends in the Evolution of the Public Web"* y Broder et al. *"Graph Structure in the web"*.

1. Definición y objetivos del estudio de la dinámica de la web

El uso global y el continuo crecimiento exponencial de la web, nos plantea una serie de desafíos a la comunidad investigadora. En particular, hay una necesidad urgente de comprender y manejar la dinámica de la web para desarrollar nuevas técnicas que hagan que la web sea tratable. ¿Y qué se entiende por dinámica de la web? Cómo cambia el uso, la topología y el contenido de la información y qué clases de modelos y técnicas estarán de acuerdo con la escala para este ritmo de crecimiento.

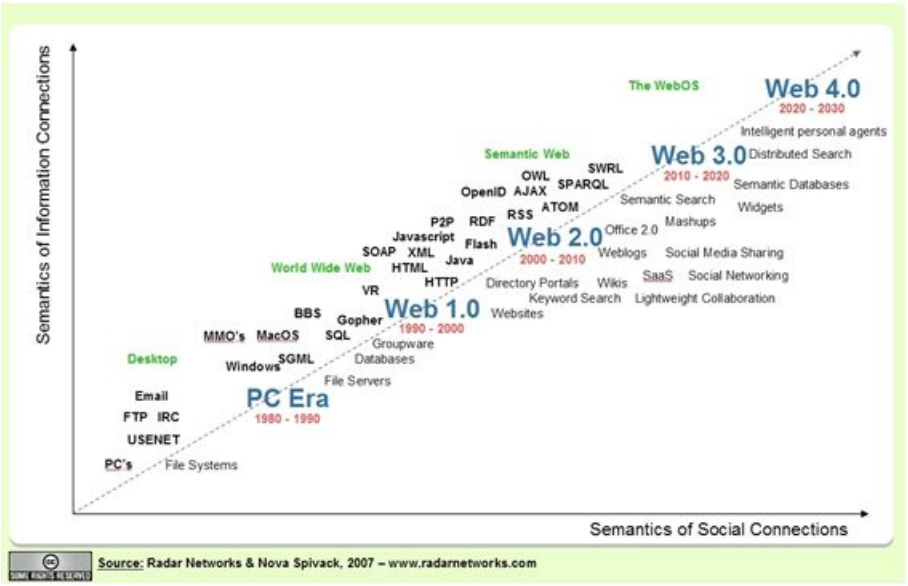


Gráfico que muestra la posible evolución futura de la web. Como se puede ver en el gráfico, la fuente es www.radarnetworks.com

2. Características de la web susceptibles de estudio

La rapidez de la ascensión de la World Wide Web del experimento alternativo al icono cultural ha sido verdaderamente notable. En menos de una década, la Web se ha extendido a casi todas las facetas de la sociedad, desde el comercio a la educación, se emplea en una variedad de usos como por ejemplo la investigación académica para una navegación rápida. Al igual que otras tecnologías de transformación que la precedieron, la Web ha generado (y consumido) grandes fortunas. La reciente "crisis de las puntocom" fue una indicación preocupante para las organizaciones de todo tipo que la naturaleza y el alcance del impacto de la Web sigue siendo inestable.

Aunque la web sigue siendo un trabajo en progreso, se ha creado mucha historia frente a un análisis significativo de las tendencias que caracterizan su evolución. Historia relativamente breve de la Web que ha sido sumergida en predicciones sobre la dirección de su desarrollo futuro, así como el papel que juega como un medio de comunicación para obtener información en forma digital. A la luz de la persistente incertidumbre que asiste a la maduración de la Web, es útil examinar algunas de las principales tendencias de la Web hasta la fecha, tanto para marcar el estado actual de evolución de la Web y para informar a nuevas predicciones sobre la evolución futura.

El artículo de *Trends in the Evolution of the Public Web* examina tres tendencias clave en el desarrollo de la Web pública - el tamaño y el crecimiento, la internacionalización, y la utilización de metadatos - sobre la base de datos de la Oficina de Investigación de la Web de OCLC Caracterización del proyecto , una iniciativa que explora cuestiones fundamentales sobre la Web y su contenido a través de una serie de muestras Web realiza anualmente desde 1998.

Las características de la web susceptibles de estudio son las siguientes:

Tamaño y crecimiento:

De acuerdo con los resultados de la encuesta más reciente del Proyecto de Caracterización Web, la Web pública, a partir de junio de 2002, contenía 3.080.000 sitios Web, o 35 por ciento de la Web en su conjunto. Los sitios públicos representaron aproximadamente 1,4 millones de páginas Web. El tamaño promedio de un sitio Web público fue de 441 páginas.

¿Es la Web pública notable en virtud de su tamaño? La respuesta rápida es no - Shapiro y Varian han estimado recientemente que el texto HTML estático en la Web fue equivalente a cerca de 1.5 millones de libros. Compararon esta cifra con el número de volúmenes en la Universidad de California en Berkeley Library (8 millones), y, teniendo en cuenta que sólo una fracción de la Web de la información puede ser considerado "útil", concluyó que "la Web no es tan impresionante como una fuente de información".

Sin embargo, la evaluación de Shapiro parece dura. La web incluye recursos digitales de muchas variedades más allá de texto plano, a menudo se combinan y recombinan en los medios de información compleja de múltiples objetos. Para evaluar el tamaño de la Web basada únicamente en texto estático se debe ignorar gran parte de la información en la Web. Por otra parte, muchos analistas Web ahora reconocen la distinción entre la "superficie Web" y la "Web profunda". Si bien esta terminología sufre de diferentes tonos de significado en contextos diferentes, la superficie de la Web se puede interpretar como la porción de la Web que es accesible a través de las tecnologías tradicionales de rastreo basado en enlace a enlace transversal de contenidos Web. Este enfoque es utilizado por los motores de búsqueda más en la generación de sus índices. La Web profunda, por el contrario, consiste en la información que es inaccesible para los rastreadores web basado en el enlace: páginas dinámicamente generadas, páginas particulares creadas en respuesta a una interacción entre el sitio y el usuario. Por ejemplo, las bases de datos en línea que generan las páginas basadas en los parámetros de consulta se considera

parte de la Web profunda. Aunque una estimación fidedigna del tamaño de la Web profunda no está disponible, se cree que es grande y creciente.

En otro estudio, Varian y Lyman estimaban que en el año 2000, la superficie de la Web representaría entre 25 a 50 terabytes de datos, basado en el supuesto de que el tamaño promedio de una página Web es de entre 10 y 20 kilobytes. Sin embargo, Varian y Lyman no hacieron ninguna distinción entre público y otros tipos de sitios Web. La combinación de sus estimaciones con los resultados de la encuesta 2000 Caracterización del proyecto Web, y suponiendo que los sitios web de todo tipo son, en promedio, el mismo tamaño en términos de número de páginas, 41 por ciento de la superficie de la Web, o entre 10 a 20 terabytes, pertenecía a la Web pública en el año 2000. En comparación, la superficie de la Web en 2002 representó el 14 hasta 28 terabytes (combinando el número de páginas del Proyecto Web de Caracterización de 2002, la encuesta dejaba un tamaño de página Web promedio de entre 10 a 20 KB).

Varian y Lyman estiman que un siti de 300 páginas, se representa con 1 MB de espacio de almacenamiento. Esto a su vez implica que a partir de junio de 2002, la información sobre la superficie de la web era más o menos equivalente al tamaño de entre 14 y 28 millones de libros. El mayor número de volúmenes, que se celebró en la Universidad de Harvard, fue un poco menos de 15 millones.

La conclusión es, sin embargo, que la Web es equivalente, o incluso supera, las colecciones de la biblioteca más grande, probablemente no.

El tamaño de la Web en su conjunto en 1996 era de alrededor de 100.000 sitios. Dos años más tarde, el primer proyecto anual de la encuesta de Caracterización Web calcula el tamaño de la Web pública en 1.5 millones de sitios. En 2000, la Web pública se había expandido a 2.9 millones de sitios, y dos años más tarde, en 2002, a más de 3 millones de sitios. En los cinco años que abarca la serie de encuestas web Caracterización del proyecto (1998 - 2002), la Web pública más del doble de tamaño.

En el estudio realizado por Edward T. O'Neill, Brian F. Lavoie y Rick Bennett, se podía observar que había un estancamiento en el crecimiento de las webs. Este estancamiento desde el 2003 hasta hoy se ha extinguido por completo y el crecimiento de Internet ha sido muy fuerte.

Distribución internacional y lenguaje.

Como su nombre indica, la World Wide Web es un recurso global de la información en el sentido de que cualquier persona, sin importar el país o idioma, es libre de hacer la información disponible en este espacio. Lo ideal sería, entonces, que el contenido de la Web se debe reflejar a la comunidad internacional en general, provenientes de fuentes en todo el mundo, y se expresa en una amplia gama de idiomas.

En 1999, el segundo año de la encuesta del Proyecto de Caracterización Web , los sitios web públicos identificados en la muestra se remontan a las entidades - individuos, organizaciones o preocupaciones de las empresas - situadas en 76 países diferentes, lo que sugiere que el contenido de la Web en ese momento era bastante inclusivo en términos de la comunidad mundial. Un examen más detenido de los datos, sin embargo, desmiente esta conclusión. De hecho, aproximadamente la mitad de todos los sitios web públicos se asociaron con las entidades ubicadas en los Estados Unidos. Ningún otro país representó más del 5 por ciento de los sitios web públicos, y sólo ocho países, además de los EE.UU., representaron más del 1 por ciento. Es evidente que, en 1999, la Web fue un espacio de información centrado en Estados Unidos.

Tres años después, poco cambió el panorama. La proporción de los sitios web públicos procedentes de fuentes de EE.UU. aumentó ligeramente en 2002, al 55 por ciento, mientras que la proporción que corresponde a los países líderes de otros permanecieron más o menos igual. En 2002, como en 1999, la

muestra contiene sitios públicos procedentes de un total de 76 países. Estos resultados sugieren que la web no está mostrando una tendencia apreciable hacia una mayor internacionalización.

Esta conclusión se refuerza cuando el idioma del contenido textual es considerada. Dado el hecho de que más de la mitad de todos los sitios públicos procedan de fuentes de EE.UU., es fácil predecir que el Inglés es el idioma más común en la Web. Pero ¿en qué medida se extiende este dominio ? ¿Cómo ha evolucionado con el tiempo?

El examen de los datos de 1999 y 2002 de la encuesta de Caracterización proporciona información sobre estas cuestiones. En 1999, 29 idiomas diferentes se identificaron en la muestra de los sitios web públicos incluidos en la encuesta, que, tomados a su valor nominal, sugiere que el contenido basado en texto de la web es bastante diversa en cuanto a los idiomas en que se expresa. Pero, como con el origen geográfico de los sitios web públicos, el total de primas de idiomas representados en la web pública exagera el grado de internacionalización alcanzado realmente. Los datos de 1999 indican que casi tres cuartas partes de todos los sitios Web públicos expresó una porción significativa de su contenido de texto en Inglés. El próximo idioma más frecuente era el alemán, que apareció en un 7 por ciento de los sitios. Sólo siete idiomas, aparte del Inglés, están representados en un 2 por ciento o más de los sitios web públicos identificados en la encuesta.

Uso de metadatos

Las bibliotecas sirven como algo más que depósitos de información. Además, la información está organizada e indexada para facilitar la búsqueda y recuperación. Una queja que a menudo se ha hecho sobre la Web es que carece de esta organización. La búsqueda se hace usando "fuerza bruta" de métodos tales como palabras clave de indexación, a menudo fuera de contexto o criterios adicionales de búsqueda. Algunas mejoras se han hecho desde los primeros días de la Web: el motor de búsqueda Google, por ejemplo, emplea algoritmos relativamente sofisticados que clasificar los resultados de búsqueda basados en los patrones de vinculación y la popularidad.

Bibliotecarios logran su organización a través de la cuidadosa preparación y mantenimiento de los datos bibliográficos - es decir, la información descriptiva sobre los recursos en sus colecciones. De manera más general, esta información descriptiva se llama metadatos, o "datos sobre datos". Un movimiento ha estado en marcha desde hace algún tiempo para introducir los metadatos en la Web, sobre todo a través de la Iniciativa de Metadatos Dublin Core. ¿Hay algún progreso importante realizado en este sentido?

Los Metadatos para los recursos Web se implementan normalmente con la etiqueta META, que puede ser utilizado por los creadores para integrar una cantidad de información que se considere relevante para describir el recurso. La etiqueta META consiste en dos componentes principales: NOMBRE, que identifica una pieza particular de metadatos (palabras clave, autor, etc) y el contenido, lo que crea una instancia, o proporciona un valor para el elemento de metadatos identificados en el atributo NAME.

Utilizando los datos de las cinco encuestas web, fue posible examinar las tendencias en el uso de metadatos en la Web pública en los últimos cinco años. El objetivo del análisis era simplemente para detectar la presencia de cualquier forma de metadatos, en ejecución usando la etiqueta META, en los sitios web públicos. El análisis de los sitios públicos recogidos en las muestras entre los años 1998 y 2002 puso de manifiesto varias características importantes sobre el uso de metadatos en la Web. En primer lugar, parece claro que el uso de metadatos va en aumento: los constantes aumentos en el porcentaje de los sitios web que contienen metadatos en la página principal (donde el uso de metadatos es la más común) se observan durante todo el período de cinco años. Incrementos similares se observaron en el porcentaje de todas las páginas Web recogerá a partir de los sitios públicos que contenían alguna forma

de metadatos.

Sin embargo debo destacar una advertencia: con la llegada de los editores más sofisticados de HTML, algunas etiquetas META se crean y se rellenan de forma automática como parte de la plantilla de documento. Esto explica cierto aumento en el uso de estas etiquetas META en sitios web públicos.

Una segunda característica de interés sobre el uso de metadatos en la Web es que, al parecer, no está cada vez más detallado. Si se asume que una etiqueta META es equivalente a un elemento de metadatos, o un pedazo de información descriptiva sobre el recurso Web, entonces es claro que, en promedio, las páginas Web que incluyen metadatos contienen alrededor de dos o tres elementos. Claramente, no hay un movimiento generalizado para incluir una descripción detallada de los recursos Web en la Web pública.

Uno de los aspectos desalentadores de las tendencias de uso de metadatos en la Web pública en los últimos cinco años es la aparente renuncia de los creadores de contenido para adoptar esquemas de metadatos formales que para describir los documentos. Por ejemplo, los metadatos Dublin Core aparecieron en sólo el 0,5 por ciento del público principal del sitio las páginas web en 1998, esa cifra aumentó casi imperceptiblemente al 0,7 por ciento en 2002. La gran mayoría de los metadatos suministrados en la Web pública es ad hoc en su creación, no estructurado por cualquier esquema de metadatos formales.

3. Ley de Zipf, "power laws" en la web

La ley de Zipf, llamada así por el profesor de lingüística de la Universidad de Harvard George Kingsley Zipf (1902-1950), es una curiosidad matemática que explica algunas de las dificultades que aparecen en las bibliotecas digitales. Supongamos que hacemos una prelación ("ranking") de las palabras que aparecen en la biblioteca de forma que asignamos el número uno a la palabra más frecuente, el dos a la segunda más frecuente, etc. Por ejemplo, en la biblioteca Miguel de Cervantes, las 10 palabras mas frecuentes y sus frecuencias de aparición f(n) son las siguientes:

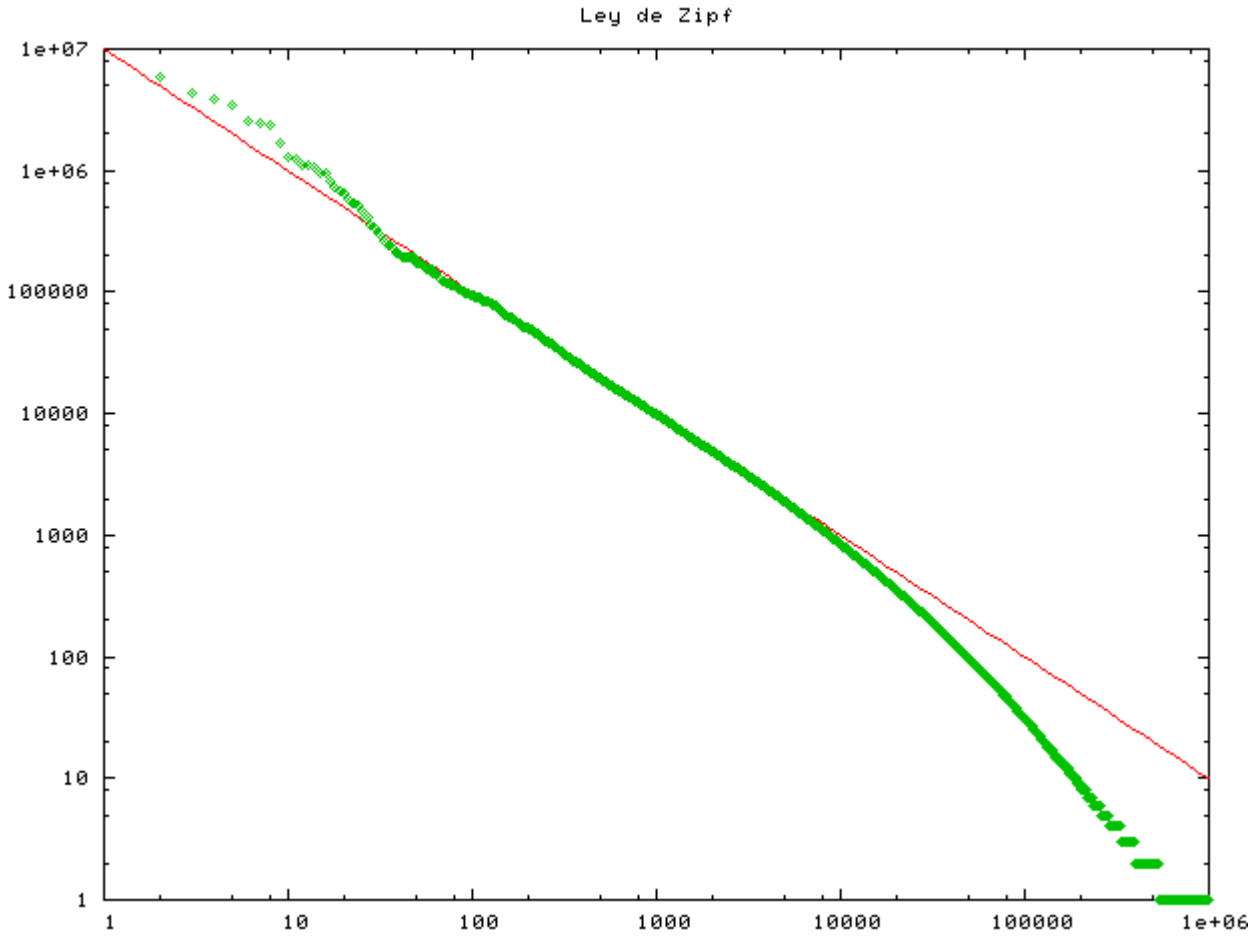
n	palabra	$f(n)$
1	de	5952871
2	que	4294496
3	y	3887331
4	la	3473934
5	en	2521954
6	el	2463429
7	a	2348470
8	los	1689770
9	se	1305932
10	no	1261456

La ley Zipf establece que el número de apariciones de una palabra es inversamente proporcional a su número de orden, es decir,

$$f(n) \simeq \frac{C}{n}$$

donde C es una constante que se fija experimentalmente.

La siguiente figura ilustra el nivel de aproximación con se cumple la ley de Zipf en la biblioteca Miguel de Cervantes (C se eligió igual a 10 millones). La línea verde representa el comportamiento ideal y los puntos rojos los valores reales.



La siguiente tabla ilustra también el nivel de aproximación para algunas palabras en particular:

n	palabra	$f(n)$	C/n
10	no	1261456	1000000
100	día	93619	100000
1000	penas	9837	10000
1000	francamente	841	1000

Pese a tratarse de un resultado aproximado, una de las virtudes de la ley de Zipf es que explica lo difícil que es construir buenos diccionarios. En primer lugar, sea cual sea el tamaño de la biblioteca, la adición de nuevos documentos añade algunas palabras nuevas. En segundo lugar, un razonamiento matemático simple nos dice que la biblioteca contiene del orden de C palabras distintas y que el número de palabras Nf con frecuencia f es aproximadamente

$$N_f \simeq \frac{C}{f(f+1)}$$

Por tanto, si construimos un diccionario que contiene todas las palabras de la biblioteca (esto es, con C entradas), aproximadamente la mitad de las palabras del diccionario (C/2) aparecen solo una vez en la biblioteca: este es, en efecto, el resultado si hacemos f = 1 en la fórmula anterior. Dicho de otra manera, si eliminamos los "hapax legomena", el tamaño del diccionario se reduce a la mitad. Adem´as, las palabras que aparecen solo dos veces (f = 2) constituyen otra parte considerable cerca del diccionario (sobre un sexto), y así sucesivamente. Es evidente que la probabilidad de cometer errores en la incorporación al diccionario de estas palabras infrecuentes es muy elevada y requiere una tarea de supervisión fabulosa. Por otro lado, es sabido que

$$\sum_{n=1}^{N} \frac{1}{n} \simeq log2N$$

Por tanto, si llamamos cobertura r del diccionario a la tasa de palabras de la biblioteca presentes en el diccionario formado por las n palabras más frecuentes, tenemos que

$$r \simeq \frac{log2n}{log2N}$$

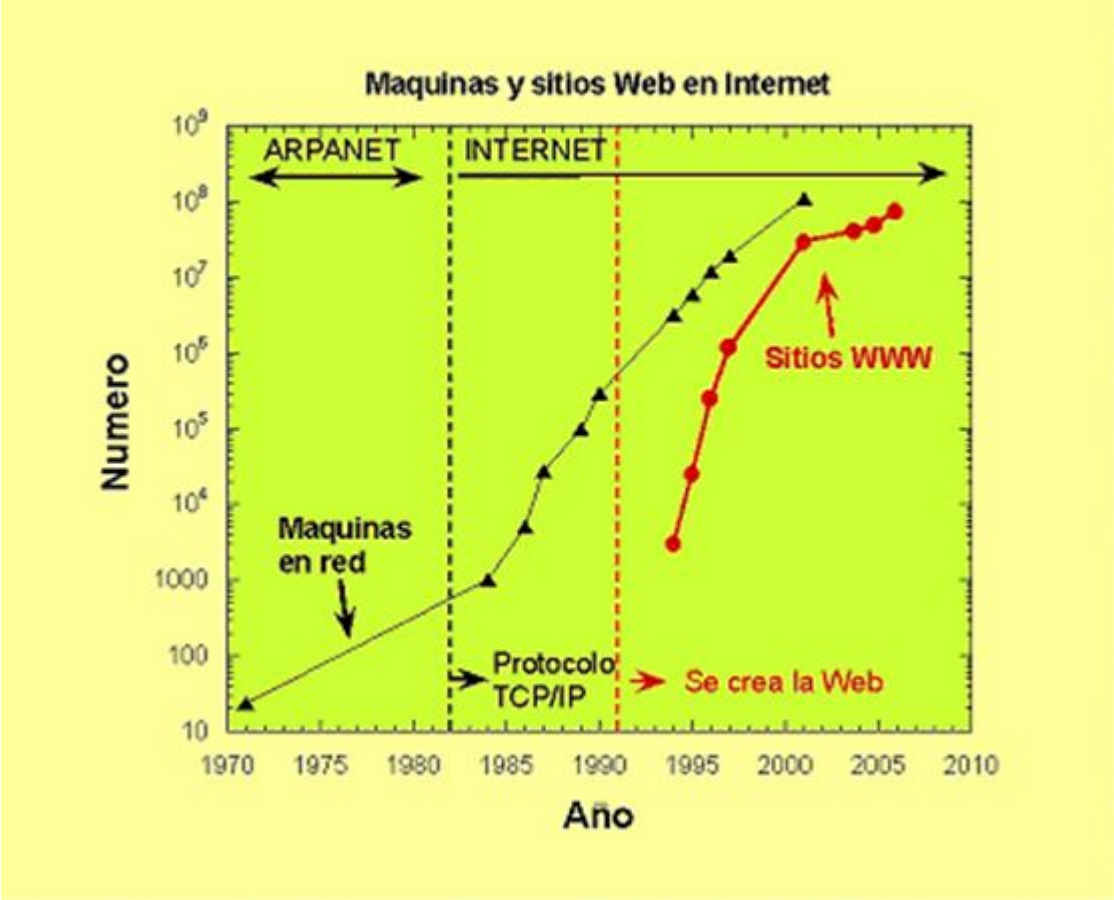
siendo N el número de palabras distintas en la biblioteca y n el número de entradas del diccionario. En nuestra biblioteca $N \simeq 1000000$ y con un diccionario con solo una décima parte de las palabras, esto es n = N/10, la cobertura se acerca al 85 %. Mejorar la cobertura en un 10 % adicional requiere incluir la mitad de las palabras (n = N/2), esto es, multiplicar por cinco el tamaño del diccionario.

4. Tamaño y tendencia de crecimiento de la web

El tamaño de la web la convierte en una fuente de información extremadamente importante, pese a que ha habido afirmaciones académicas como la de Shapiro y Varian que han tratado de restarle importancia debido a dos causas: El texto contenido en la Web era menor que el de su biblioteca en 2002 y una segunda afirmación sobre el hecho de que sólo un porcentaje de la información que hay en la red es "útil". Debido al ritmo de crecimiento que tiene hoy la red, la cantidad de información

disponible es muy grande. Y no sólo se reduce a texto plano tal y como ellos afirmaban. Al descartar los objetos multimedia no tenían en cuenta una gran parte de la información de la web. De todas formas, no se puede decir taxativamente que la Web sobrepase o sea equivalente, a las mayores colecciones de libros, ya que un porcentaje significativo de la Web superficial está formada por "Encabezado de formato", por ejemplo etiquetas XML o HTML. Tampoco puede descartarse fácilmente la idea de que una porción significativa de la información de la Web no es útil.

Lo más interesante sobre el tamaño de la Web es cuán rápidamente se levantó desde proporciones relativamente insignificantes hasta una escala al menos comparable a las colecciones de librerías de investigación:



Nos podemos dar cuenta de que si bien el ritmo sigue siendo de crecimiento, éste se produce a un ritmo mucho menor. Esto es debido a que la Web ya no es una nueva tecnología. Aquellos que desean establecer una presencia en la Web prbablemente ya lo han hecho. En este sentido, la prisa por "Ponerse online" que se verificaba durante los primeros años de la Web ha dejado paso a un deseo de refinar y desarrollar los sitios web existentes. En el enlace [6] hay una estimación más moderna del número de páginas web en la fecha de realización de este resumen

5. Web pública y web oculta

Hay estimaciones que calculan en 500 veces más grande que el total de la información indizada por los buscadores, la información que permanece invisible en la World Wide Web, lo que se ha denominado el inmenso océano de la Internet profunda.

Aunque los buscadores generalistas no suelen indizar archivos no textuales, sí existen una serie de buscadores especializados que indizan imágenes, vídeo, audio, archivos pdf, archivos comprimidos o ejecutables. Sin embargo, muchísimos datos quedan fuera de los buscadores tradicionales, ya sean estos generalistas o especializados, puesto que indizar cierto tipo de informaciones contenidas en enormes bases de datos numéricas o textuales, exige gastar muchos recursos y resulta muy costoso para los buscadores almacenar en sus bases de datos este tipo de formatos. Por otro lado, los buscadores tampoco indizan muchos de los datos que se generan de forma dinámica en tiempo real, puesto que se convierten en obsoletos en un brevísimo lapso de tiempo y no merece gastar recursos en informaciones tan fugaces; y a esto se une que muchas de estas bases de datos dinámicas han de rastrearse desde su propia ubicación o sitio web, y con sus propias herramientas de búsqueda personalizadas, puesto que precisan de pasarelas o contraseñas especiales para acceder a ellas. Si a esto unimos las páginas sin conexión o enlaces aparentes, vemos que una enorme masa de información no es accesible desde los principales buscadores existentes en la World Wide Web. A toda esta gran masa de información es a la que se ha denominado Internet oculta.

Según el estudio How much Information? 2003, realizado por Peter Lyman y Hal R. Varian de la School of Information Management and Systems de la Universidad de California, Berkeley, la cantidad de información de la Web navegable o visible es de 147 terabytes, mientras que la Web invisible es de 91.850 terabytes.

¿Qué información es la que permanece invisible? Toda aquella información almacenada en bases de datos, material de archivo y herramientas interactivas tales como diccionarios o calculadoras, páginas dinámicas construidas con tecnologías Flash, ASP, PHP, etc. Estos recursos son embebidos dentro de miles de sitios web individuales y no son "visibles" para los motores de búsqueda tradicionales. Para acceder a todo ese incalculable acervo de información sólo podemos interrogar a las bases de datos directa e individualmente a través de sus propios formularios de búsqueda, puesto que las páginas indizables por los motores de búsqueda no dan cuenta de los recursos en ellas disponibles. Lo que está claro es que nadie tiene acceso completo a todo Internet ya que no sólo existen áreas concretas de la red que son inaccesibles a la mayor parte de los internautas, sino también determinados contenidos que permanecen invisibles.

Ricardo Fornas Carrasco en La cara oculta de Internet establece 3 tipos distintos de Internet:

- Internet global: Definiremos ésta como aquella Red de información libre y gratuita que es accesible teóricamente mediante la interconexión de ordenadores. La forma de acceso se realiza mediante programas navegadores, Chats, mensajería o intercambio de protocolos (FTP, P2P).

- Internet invisible: Responde a todos aquellos contenidos de información que están disponibles en Internet pero que únicamente son accesibles a través de páginas generadas dinámicamente tras realizar una consulta en una base de datos. Esta particular naturaleza les hace inaccesibles a los procesos habituales de recuperación de la información que realizan buscadores, directorios y agentes de búsqueda. Pero podemos acceder a las mismas mediante nuestras habituales herramientas de navegación, correo, etcétera. La única condición es saber exactamente la dirección de acceso (URL o FTP)

- Internet oscuro: Se define como los servidores o host que son totalmente inaccesibles desde nuestro ordenador. Según un estudio de la compañía Arbors Networks esta situación sucede en el 5 % de los contenidos globales de la Red. La causa principal (78 % de los casos) se debe a zonas restringidas con fines de seguridad nacional y militar. No olvidemos que Internet es un invento militar. El porcentaje restante, (22 %) obedece a otros motivos: configuración incorrecta de routers, servicios de cortafuegos y protección, servidores inactivos y finalmente "secuestro" de servidores para utilización ilegal.

Al igual que Internet invisible, la denominada Web invisible contiene un gran número de fuentes de información que no pueden buscarse porque su contenido no ha sido indexado ni puede serlo por los principales buscadores. Aun cuando recuperemos un sitio que contenga una base de datos, es improbable que el buscador conduzca a la base de datos misma, puesto que requiere que se navegue por el sitio web para encontrarla. Así pues, la Web invisible está constituida por toda esa información accesible vía web, pero a la que no es posible llegar mediante una consulta a los buscadores tradicionales.

6. Idiomas en la web

Idealmente, el contenido de la Web debería reflejar toda la comunidad internacional, originándose de fuentes de todo el mundo y expresadas en un amplio rango de lenguas. La realidad es que, los principales responsables de este esfuerzo de crecimiento de Internet son Estados unidos, Alemania, China, Corea del Sur y Japón, estando la inmensa mayoría de esto sitios en inglés.

La red de redes se utliliza fundamentalmente para buscar información. Existen posibilidades infinitas en ella.

Por sus propias características y las posibilidades que ofrece, Internet es un medio de comunicación enormemente potente y en constante crecimiento alrededor del globo. En todos los países y culturas se la utiliza para intercambiar datos e informarse.

Aunque la supremacía del inglés en Internet es abrumadora, nos encontramos ante un medio que, casi por definición, ha de ser también multilingüe, y es muy común encontrarse con botones o marcas que nos permiten elegir el idioma en el queremos leer un texto. Casi la totalidad de los buscadores ofrece la opción de traducir la página que estamos viendo en el idioma que uno desea y afortunadamente existen potenes traductores gratuitos que pueden ser usados desde la red.

¿Cuáles son los idiomas mas utilizados en Internet?

Basándome en datos publicados por Internet World Stat (`http://www.internetworldstats.com/`), que es una compañia internacional dedicada a informar diariamente sobre la utilización de Internet en el mundo, brindando datos y estadísticas por países y regiones . Según los datos relevados a nivel mundial, el 36 % de los internautas proviene de Asia (con 418 millones de usuarios), el 28 % de Europa (con 322 millones de usuarios), el 20 % de Norteamérica (con 233 millones de usuarios) y el 9 % de Latinoamérica (con 110 millones de usuarios). El 7 % restante se reparte entre Oceanía, Medio Oriente y Africa.

Los idiomas mencionados en la lista representan el 80 % de los idiomas de la red, ya que el 20 % (alrededor de unos 200 millones) utilizan otros idiomas no mencionados.

En cuanto al idioma Español es notable el crecimiento que el mercado hispano tuvo dentro de la Web. Se estima que solo el 25 % de los hispanoparlantes tiene acceso a Internet y se cree que este

número irá en ascenso. La mayor parte de internautas que utiliza el idioma Español proviene de los Estados Unidos.

Estos datos son sumamente interesantes para los Webmasters (creadores de páginas web) ya que sabrán a quiénes y en qué idioma orientar sus contenidos para llegar a un número mayor de personas. Los datos fueron publicados a finales de marzo de 2007. A continuación detallo una lista de los 10 idiomas que son mas utilizados en Internet detallando la cantidad de internautas que habla cada lengua.

1. Inglés 329 millones de usuarios.
2. Chino 159 millones de usuarios.
3. Español 89 millones de usuarios.
4. Japonés 86 millones de usuarios.
5. Alemán 59 millones de usuarios.
6. Francés 56 millones de usuarios.
7. Portugués 40 millones de usuarios.
8. Coreano 34 millones de usuarios.
9. Italiano 31 millones de usuarios.
10. Arabe 28 millones de usuarios.

7. Dominios en la web

Los nombres de dominio son la traducción para las personas de las direcciones IP, las cuales son útiles sólo para los ordenadores.

Así, por ejemplo, google.es es un nombre de dominio. Como se puede ver, los nombres de dominio son palabras separadas por puntos, en vez de números en el caso de las direcciones IP. Estas palabras pueden darnos idea del ordenador al que nos estamos refiriendo. Si se sabe un poco más sobre nombres de dominio, con sólo ver google.es podremos concluir que "Una empresa de España que da cierta información por Internet es Google. En Estados Unidos la última palabra del nombre del dominio representa qué tipo de organización posee el ordenador al que nos referimos

- com: Empresas
- edu: Instituciones de carácter educativo, mayormente universidades
- org: Organizaciones no gubernamentales
- gov: Entidades del gobierno
- mil: Instalaciones militares
- info: Organizaciones que ofrecen información
- tv: Cadenas de televisión

En el resto de los países, la última palabra indica el país:

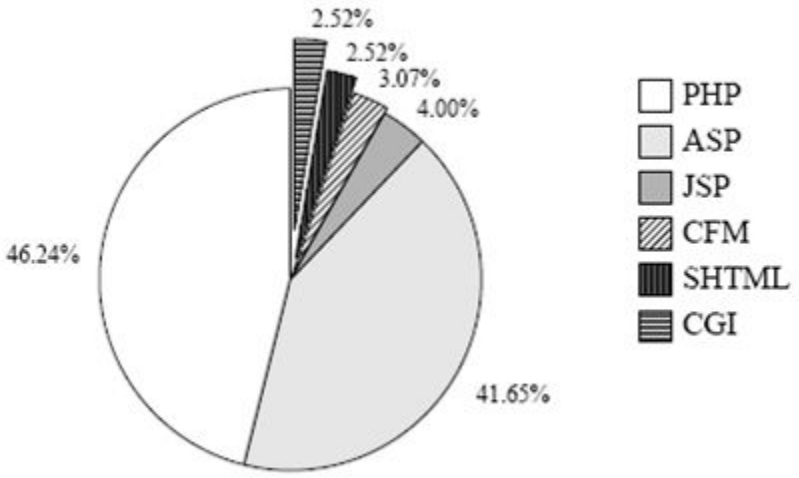
- es: España
- fr: Francia
- uk: Reino unido
- it: Italia
- jp: Japón
- au: Australia
- ch: Suiza Para una lista más exhaustiva, consultar el enlace [9].

Una de las listas más completas y actualizadas de los dominios que existen actualmente se encuentra en el siguiente enlace:
`http://www.iana.org/domains/root/db/`

8. Estudios sobre la web española

Las principales conclusiones que se desprenden del estudio de la web española realizado son que:

- Una gran cantidad de sitios no utilizan el dominio de primer nombre correspondiente al país .es, prefiriendo .com o .org
- Por eso mismo, el domino .es se ha visto libre de malas prácticas
- Una gran parte de la información disponible en la web de España es generada por universidades o entidades de gobierno.
- La prensa escrita también tiene una participación importante, tanto en referencias como en número de páginas
- Las propiedades estadísticas de la muestra son muy similares a las de otras muestras, con lo que se indica que la muestra puede ser usada para estudios que sean al menos parcialmente extrapolables a la red global
- Un 1 % de las páginas de la web son enlaces a ficheros que no son HTML. Si bien parece un número pequeño, son unos 200000 documentos. Los formatos de texto plano y pdf son los más usados
- La proporción de sitios que sólo constan de una página, sin ningún enlace, es cercana al 30 %
- El siguiente gráfico nos da la información de las herramientas usadas para crear páginas dinámicas, siendo las principales el PHP y el ASP



9. Áreas de investigación relacionadas

La dinámica de la web está directamente relacionada con las siguientes áreas de investigación, que ya hemos tratado a lo largo de todo el curso en los distintos trabajos realizados:

1. Minería de datos : Se puede decir que la minería de datos (DM, Data Mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

2. Teoría de grafos: En matemáticas y en ciencias de la computación, la teoría de grafos (también llamada teoría de las gráficas) estudia las propiedades de los grafos (también llamadas gráficas). Un grafo es un conjunto, no vacío, de objetos llamados vértices (o nodos) y una selección de pares de vértices, llamados aristas (edges en inglés) que pueden ser orientados o no. Típicamente, un grafo se representa mediante una serie de puntos (los vértices) conectados por líneas (las aristas).

3. Recuperación de la información: La Búsqueda y Recuperación de Información, llamada en inglés Information Search and Retrieval (ISR), es la ciencia de la búsqueda de información en documentos electrónicos y cualquier tipo de colección documental digital, encargada de la búsqueda dentro de éstos mismos, búsqueda de metadatos que describan documentos, o también la búsqueda en bases de datos relacionales, ya sea a través de internet, intranet, y como objetivo realiza la recuperación en textos, imágenes, sonido o datos de otras características, de manera pertinente y relevante. La recuperación de información es un estudio interdisciplinario. Cubre

tantas disciplinas que eso genera normalmente un conocimiento parcial desde tan solo una u otra perspectiva. Algunas de las disciplinas que se ocupan de estos estudios son la psicología cognitiva, la arquitectura de la información, diseño de la información, inteligencia artificial, lingüística, semiótica, informática, biblioteconomía, archivística y documentación. Para alcanzar su objetivo de recuperación se sustenta en los sistemas de información, y al ser de carácter multidisciplinario intervienen bibliotecólogos para determinar criterio de búsqueda, la relevancia y pertinencia de los términos, en conjunto con la informática.

10. Conferencias internacionales

Algunas de las conferencias internacionales que abordan el tema de la búsqueda web, indexación y métodos de indexación son las siguientes:

- International World Wide Web Conference(IW3C2).
- International journal of Computer Networks & Communications (IJCNC)
- International Conference on Internet and Web Engineering
- Interlink Web Design Conference
- International Conference on Web Intelligence, Mining and Semantics
- International Conference on Web-based Learning (ICWL 2010)
- International Conference on Machine Learning (ICML97)
- International Conference on Autonomous Agents (Agents '98)
- International Conference on Web Information Systems and Technologies

Referencias

[1] Shapiro, C. and H. Varian (1998). Information Rules: A Strategic Guide to the Network Economy, (Harvard Business School Press, Cambridge)

[2] Statistics from Gray, M., "Web Growth Summary". Available at <http://www.mit.edu/people/mkgray/net/web-growth-summary.html>.

[3] Mariano, G. (2002). "The Incredible Shrinking Internet" ZDNet UK News. Available at <http://news.zdnet.co.uk/story/0,,t269-s2101890,00.html>.

[4] WorldCat statistics were obtained from the OCLC Annual Report 2000/2001. The report is available at <http://www.oclc.org/about/annualreport/2001.pdf>.

[5] http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

[6] M. Levene and A. Poulovassilis. Web Dynamics. Software Focus, 2, (2001), 60-67.

[7] Ricardo Baeza-Yates, Bárbara J. Poblete y Felipe Saint-Jean. Evolución de la Web Chilena. Centro de Investigación de la Web, 2003.

[8] Edward T. O'Neill, Brian F. Lavoie, Rick Bennett. Trends in the Evolution of the Public Web (1998 - 2002). D-Lib Magazine,Volume 9 Number 4, April 2003.

[9] Broder et al. Graph Structure in the web. Proc.WWW9, 2000.

[10] Baeza-Yates y C. Castillo. Caracterizando la Web Chilena. Encuen- tro Chileno de Ciencias de la Computación, año 2000. Disponible en http://www.todocl.cl/stats.phtml

[11] L.A. Adamic. The small world web. In Proceedings of European Conference on Research and Advanced Technology for Digital Libraries, pages 443–452, Paris, 1999.

[12] S. Abiteboul, P. Buneman, and D. Suciu. Data on the Web: From Relations to Semistrucutred Data and XML. Morgan-Kaufmann, San Francisco, Ca., 2000

[13] J. Borges and M. Levene. A fine grained heuristic to capture web navigation patterns. SIGKDD Explorations, 2:40–50, 2000.

[14] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of International World Wide Web Conference, pages 107–117, Brisbane, 1998.

[15] P. Brusilovsky, A. Kobsa, and J. Vassileva, editors. Adaptive Hypertext and Hypermedia. Kluwer, Dordrecht, 1998.

[16] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In Proceedings of International World Wide Web Conference, pages 1623–1640, Montreal, 1999.

[17] J. Chen, D. DeWitt, F. Tian, and Y. Wang. Niagaracq: a scalable continuous query system for internet databases. In Proc. 2000 ACM SIGMOD Int. Conf.on Management of Data, pages 379–390, 2000.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The web as a graph. In Proceedings of ACM Symposium on Principles of Database Systems, pages 1–10, Dallas, Tx., 2000.

[19] D. Milojicic. Trend wars - Mobile agent applications. IEEE Concurreny, 7:80–90, 1999