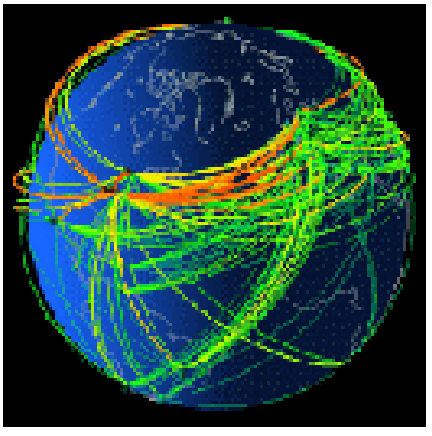
aguja en un pajar.

ABSTRACT

We present the main characteristics of the Web, including its size, structure, and languages. We describe the main tools currently available to search the Web and illustrate the technology behind them through a few examples. The best use of the wealth of information available on the Web will depend on whether these technologies will evolve as fast as the Web grows.

Searching in the World Wide Web may be harder than finding a needle in a haystack..

Wide Web (Web de ahora en adelante, aunque no queda claro si es femenino o masculino)? Nadie sabe. Crece más rápido que la capacidad de ella misma



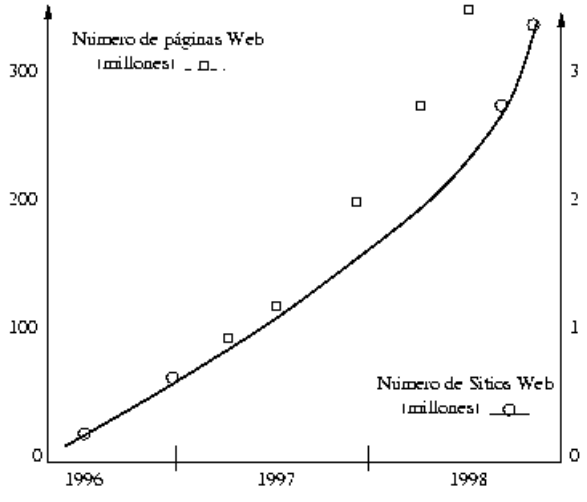
caracterizamos algunos aspectos de la Web, incluyendo su impacto en Iberoamérica y el castellano, concluyendo con las formas que existen para buscar en ella y cómo funcionan los buscadores de información en la Web.

Estructura y Visibilidad

¿Cuántas referencias tiene una página HTML? (HTML es un acrónimo para Hyper Text Markup Language, el lenguaje usada para estructurar páginas

Web) Más del 75% de las páginas tiene al menos una referencia, y en promedio cada una tiene entre 5 y 15 referencias. La mayoría de estas referencias son a páginas en el mismo servidor. De hecho, la conectividad entre sitios distintos no es muy buena. En particular, la mayoría de las páginas no son referenciadas por nadie y las que sí son referenciadas, lo son por

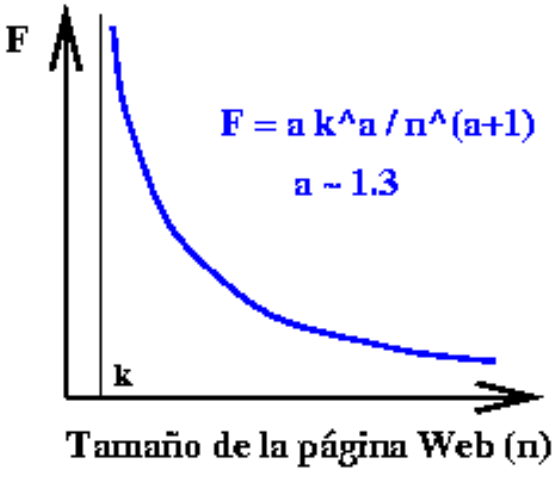
Considerando sólo referencias externas (entre sitios distintos), más del 80% de las páginas tienen menos de 10 referencias a ella. Otros sitios son muy populares, teniendo decenas de miles de referencias a ellos (por ejemplo en 1995 los top-5 eran Univ. of Illinois at Urbana- Champaign, National Institute of Health, Mass. Inst. of Tech., NASA y Carnegie Mellon Univ. [4]). Si contamos sitios que referencian a sitios, aparecen Yahoo!, Microsoft y Netscape entre los 5 primeros. Por otro lado, hay algunos sitios que no son referenciados por nadie (están porque fueron incluidos mediante el envío directo de una dirección Web a Yahoo u otros buscadores, pero que realmente son islas dentro de la Web). En este mismo sentido, las páginas personales también se pueden considerar como entes aislados en la mayoría de los casos. Asi mismo, la mayoría de los sitios (80%) no tiene ninguna referencia hacia páginas en otros servidores. Esto significa que una minoría de los servidores mantiene toda la carga navegacional de la red. En particular hay sitios que tienen miles de punteros externos que son los que al final engloban la Web, siendo obviamente el mayor de todos ellos Yahoo!. Estadísticas recientes indican que el 1% de los servidores contienen aproximadamente el 50% del volumen de datos de la Web, que se estima es de alrededor de 320 millones de páginas a mediados de 1998 [3]. El siguiente gráfico muestra el crecimiento aproximado en el último tiempo del número de servidores y páginas Web.



Tamaños y características

¿Cómo es una página Web promedio? Una página de texto promedio tiene alrededor de 5 kilobytes (poco menos de mil palabras). Si agregamos audio o video, este promedio aumenta. De hecho la distribución de tamaños se dice que es de ``cola pesada", como por ejemplo la distribución de Pareto (ver figura siguiente). En otras palabras, aunque la mayoría de los archivos son pequeños, existe un número no despreciable de archivos grandes; y hasta 50

llegando a varias decenas de megabytes, tenemos archivos de video. Los formatos más populares (en base a la extensión del nombre de archivo) son HTML, GIF, TXT, PS y JPG, en ese orden.



¿Cómo es una página HTML? Alrededor de la mitad de ellas no tiene ninguna imagen. Un 30% no tiene más de dos imágenes y su tamaño promedio es de 14Kb. Por otra parte hay un porcentaje no despreciable (mayor al 10%) de páginas con más de 10 imágenes. La razón es que son imágenes tipográficas, como por ejemplo puntos rojos, líneas de separación de color, etc. La mayoría de las páginas usan HTML simple. Sólo un porcentaje pequeño siguen todas las normas y otro porcentaje mayor (alrededor del 10%) son sólo texto. Finalmente, la calidad del texto deja mucho que desear, pues hay errores de tipeo, errores de OCR, etc. Más aún, la información contenida puede estar obsoleta, puede ser falsa o engañosa. Hay que tener esto en mente cuando usamos una página Web como fuente de información.

Funredes,
desde 1996 a la fecha la razón francés/castellano ha pasado de 2.4 a 1.1, por lo que en el año 2000 el castellano debiera ocupar ya el cuarto lugar.

Son dos las maneras más usadas para buscar. Podemos usar catálogos similares a las páginas amarillas telefónicas como Yahoo!. Estos catálogos son taxonomías jerárquicas que intentan clasificar los distintos temas o áreas del conocimiento. La ventaja principal de este método es que si encontramos

algo, seguramente será útil. Las desventajas son que la clasificación muchas veces no es suficientemente especializada y no todo lo que existe en la Web está clasificado. De hecho, la Web crece más rápido que cualquier catálogo. Los esfuerzos para realizar esto de forma automática datan de los comienzos de la inteligencia artificial en los años 60. Sin embargo hasta hoy, el procesamiento de lenguaje natural para extraer términos relevantes de un documento no es 100% efectivo.

La segunda técnica es usar una máquina de búsqueda (*search engine*) como AltaVista, Lycos o Infoseek, que usan el paradigma de recuperación en texto completo. Es decir, todas las palabras de un documento se almacenan en un índice para su posterior recuperación. Más adelante hablaremos de los desafíos técnicos para crear este índice. Un problema adicional es que el recorrer la Web actualizando y agregando nuevas páginas, es una tarea que no termina nunca y que además tampoco puede mantenerse vigente con el crecimiento continuo de la Web. Aunque las búsquedas en estas máquinas son efectivas en muchos casos, en otros son un total desastre. El problema es que las palabras no capturan toda la semántica de un documento. Hay mucha información contextual o implícita que no está escrita, pero que entendemos cuando leemos.

El siguiente ejemplo ilustra los problemas de buscar en la Web. Supongamos que queremos encontrar a qué velocidad corre un jaguar buscando las siguientes palabras: **jaguar speed** (queramos o no, el idioma más usado en la Web es inglés y tal vez tengamos que convertir millas por hora a kilómetros por hora). El resultado en AltaVista es un montón de páginas acerca del auto Jaguar, un juego de video para Atari, un equipo de fútbol americano, un servidor de redes locales, etc. ¡La primera página acerca del animal está en el lugar 183 y es una fábula! Si intentamos eliminar los documentos acerca del modelo de auto, igual encontraremos páginas acerca de **car**, ni **auto**. Tratemos **jaguar speed +cat**, que indica que la palabra **cat** (felino) debe estar en el documento. Los dos primeros resultados son acerca de los clanes Nova Cat and Smoke Jaguar, luego, la empresa LMG, seguido de automóviles finos. La número 25 es la primera con información de jaguares, pero tampoco tiene lo que necesitamos. Si miramos en Yahoo!, podemos buscar en **Science: Biology:Zoology:Animals:Cats: Wild_Cats** y en **Science:Biology:Animal_Behavior**, pero en ninguno en_ contramos una página acerca de jaguares.

Es decir, las máquinas de búsqueda todavía devuelven demasiada basura para poder encontrar la aguja mientras los catálogos no tienen la profundidad y volumen suficiente para clasificarla. El problema de ordenar documentos en base a palabras como hace AltaVista no se puede resolver bien con tan poca información (dos palabras) y adolece de la misma dificultad intrínseca de la clasificación automática. Sería más efectivo tratar de realizar

Yahoo! debieran entregar caminos en la jerarquía para asegurarnos que estamos recuperando del tema de nuestro interés. *Moraleja*: si quiere algo específico, mire una enciclopedia, para eso se crearon. Por otro lado, si no sabe exactamente lo que quiere, use una máquina de búsqueda y vaya modificando su consulta de acuerdo a los documentos que recupere y sean relevantes. O si está interesado en un tema amplio, vaya a Yahoo!. Allí encontrará buenos lugares donde comenzar a navegar.

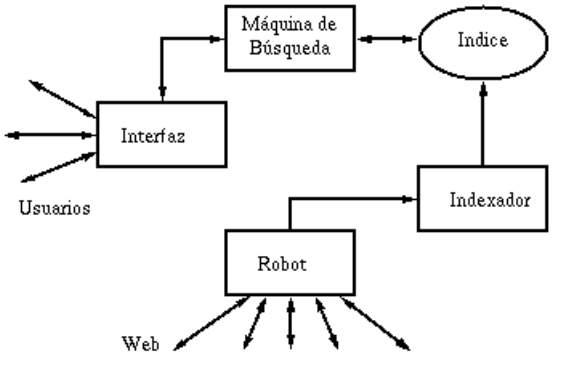
Recursos en Castellano

Si queremos buscar información en castellano, hay varias alternativas. La más simple es usar un buscador estándar, por ejemplo Altavista (que actualmente es el de mayor cobertura) y usar palabras en castellano (que no existan en otro idioma). Algunos buscadores también permiten especificar el idioma o el área geográfica. También Yahoo! tiene ahora un directorio en castellano de datos en esta lengua, con páginas específicas de 6 países, entre

Por otra parte, hay otros buscadores especializados. Por ejemplo, en España hay más de 35 de ellos, tales como Elcano, Lycos España, Ole, etc. Algunas direcciones útiles están en Argentina e Inglaterra. En Chile un buen directorio es La Brújula.

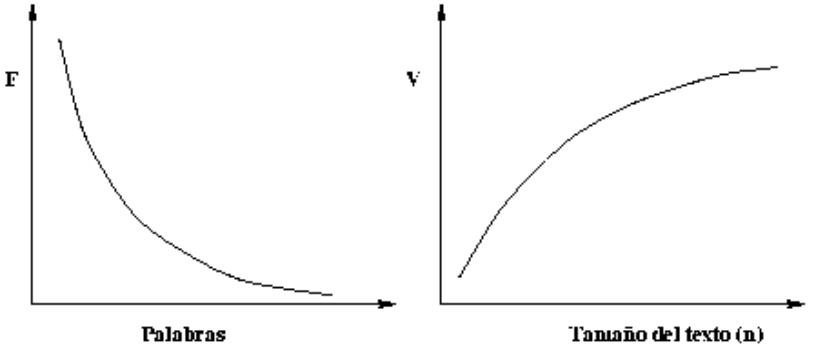
Indexando la Web

Queda claro que para extender un directorio como Yahoo! se necesitan expertos que clasifiquen nuevas páginas que en general son informadas por los propios interesados. Por otra parte, indexar toda la Web implica el uso de programas llamados *crawler*, *robot, wanderer*, etc. que recorren la Web y recopilan páginas nuevas o actualizadas. La arquitectura típica de un buscador (ver figura siguiente) incluye el indexador y el robot. A continuación hablamos de cómo crear un índice de toda la Web. 4



Nadie conoce el volumen actual de la Web. Estimemos por debajo la cantidad de texto. Si cada página tiene 5Kb y hay como 300 millones de páginas, estamos hablando de más de 1.5 Tb de texto solamente. Esta es una estimación conservadora y por supuesto el volumen total es mayor. Indices como AltaVista mantienen todas las palabras distintas ordenadas y para cada palabra la lista de páginas Web donde aparecen. Esta estructura de datos se llama archivo invertido.

texto, sino que crece en forma sublineal (Ley de Heaps). Esto se debe a que el vocabulario es finito. Por otra parte, la frecuencia de las palabras sigue una variante de la Ley de Zipf. Esta ley experimental indica que la *j*-ésima palabra más frecuente aparece una cantidad de veces proporcional al inverso de *j*. Actualmente esta distribución es más sesgada y se aproxima más al inverso del cuadrado de *j*. Es decir, hay un conjunto pequeño de palabras muy frecuentes y muchas que aparecen muy pocas veces o sólo una vez (sea cual sea el idioma usado). Estas distribuciones se presentan a continuación.



Usando distintas técnicas, el tamaño de un archivo invertido puede reducirse a un 20% del tamaño del texto, es decir al menos unos 300Gb (en la realidad es mucho más). Estos índices se pueden reducir usando particiones lógicas en vez de documentos (por ejemplo, poniendo muchas páginas pequeñas en un mismo grupo). Usando búsqueda binaria en las palabras ordenadas, podemos encontrar todos los documentos en que aparece en menos de un segundo. Dependiendo del sistema de búsqueda, estos documentos serán ordenados usando distintos criterios y heurísticas, con el objeto de indicar al usuario cuál es el documento más relevante (esto funciona muchas veces, pero otras no). Otro problema debido al volumen de datos es que la cantidad de documentos resultantes es del orden de miles, por lo cual es necesario usar paradigmas visuales para poder manipularlos. Por ejemplo, el índice de AltaVista, que es el más grande y registra sobre 100 millones de

Altavista: http://www.altavista.com/
Alis technologies: http://babel.alis.com:8080/palmares.html
Buscadores de España:
http://www.netmasters.co.uk/european_search_engines/page42.html
Excite: http://www.excite.com/
Funredes: http://funredes.org/
HotBot: http://www.hotbot.com/
Infoseek: http://www.infoseek.com/
La Brújula: http://www.brujula.cl/
Lycos: http://www.lycos.com/

Meta Miner: http://www.miner.com/
Microsoft: http://www.microsoft.com/
Netscape: http://www.netscape.com/
NorthernLight: http://www.northernlight.com/
Search Broker: http://debussy.cs.arizona.edu/sb/
Yahoo: http://www.yahoo.com/
Yahoo en español: http://espanol.yahoo.com/

Referencias

[1] Marc Abrams (editor), World Wide Web: Beyond the Basics, Prentice Hall, 1998: http://ei.cs.vt.edu/~wwwbtb/hardcopy/book/

[2] Ricardo Baeza-Yates y Berthier Ribeiro-Neto, Modern Information Retrieval (Capítulo 13: Searching the Web), Addison-Wesley-Longman, Wokingham, Inglaterra, Marzo 1999.

[3] K. Bharat y A.Z. Broder, A Technique to measuring the relative size and overlap of public Web search engines, 7th WWW Conference, Brisbane, Australia, 379-388: http://www.research.digital.com/SRC/whatsnew/semchart.html

[4] Tim Bray, Measuring the Web, Fifth International World Wide Web Conference, Paris, Mayo 1996: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

[5] Martin Dodge, The Geography of Cyberspace Directory: Main Page, 1997: http://www.geog.ucl.ac.uk/casa/martin/geography_of_cyberspace.html

[6] Netcraft Web Server Survey, 1998: http://www.netcraft.com/Survey/

[7] NetSizer: Main Page, 1998: http://www.netsizer.com/

[8] Network Wizards, Internet Domain Survey, 1998: http://www.nw.com/

[9] Greg Notess, Search Engines Showdown: Main Page, 1998: http://www.notess.com/search/

[10] OCLC, Study of Web Characteristics, 1998: http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm

[11] Danny Sullivan, Search Engine Watch: Main Page, 1997: http://www.searchenginewatch.com/

Ricardo Baeza Yates es Ph.D. en Computer Science (Univ. of Waterloo, Canadá,1989), Magister en Ing. Eléctrica (1986) y Cs. de la Computación (1985) de la Univ. de Chile; e Ingeniero Civil Eléctrico de la misma universidad. Actualmente es Profesor Titular en el Depto. de Cs. de la Computación de la Univ. de Chile y sus áreas de investigación son algoritmos, bases de datos documentales y visualización. Es co-autor de un Handbook de algoritmos (Addison-Wesley, 1991) y co-editor de un libro en recuperación de la información (Prentice-Hall, 1992),además de numerosas actas de congresos, publicaciones internacionales y nacionales. Es el presidente de la Sociedad Chilena de Ciencia de la Computación (1997-1998), cargo que ocupó tambien desde el año 1993 al 1995.

Si tiene preguntas o sugerencias, envíe e-mail a *rbaeza@dcc.uchile.cl*