Tecnología

Desenredando La Madeja

Ricardo Baeza Yates
Universidad de Chile

<rbaeza@dcc.uchile.cl >

Resumen: *en este artículo analizamos el problema de buscar información en la web. Para ello presentamos las características principales de la web, incluyendo su tamaño, estructura e idiomas, con un énfasis en Iberoamérica. Luego, a través de varios ejemplos, describimos las herramientas que existen actualmente para buscar en la web y la tecnología que está detrás de ellas. El buen uso de la información que existe en la web dependerá de si estas tecnologías pueden evolucionar tan rápido como crece la web. Buscar en World Wide Web puede ser más difícil que encontrar una aguja en un pajar.*

1. Introducción

¿Qué estructura tiene la telaraña mundial de computadores o *World Wide Web*? (la *web* de ahora en adelante, aunque no me queda claro si es femenino o masculino) Nadie sabe. Crece más rápido que la capacidad de ella misma para detectar sus cambios. Sus conexiones son dinámicas y muchas de ellas quedan obsoletas sin ser nunca actualizadas. El contenido de la *web* es hoy de varios terabytes (un terabyte o TB es un billón de megabytes) de texto, imágenes, audio y video. Para aprovechar esta gran base de datos no estructurada, es importante poder buscar información en ella, adaptándose al crecimiento continuo de la *web*.

Al igual que Internet, la red de computadores que interconecta el globo, que ya sobrepasó los 77 millones de computadores conectados en más de 220 países [8], los servidores de *web* también crecen en forma exponencial desde 1993 (un servidor *web* es el software que administra un sitio *web*). Lamentablemente nadie sabe su número exacto, pues no es posible a partir de un nombre de dominio saber si es o no un servidor *web* (la mayoría comienza con *www*, pero muchos lugares no siguen esta convención). Además un mismo computador puede manejar distintos servidores y también existen servidores virtuales (un mismo servidor *web* puede manejar lógicamente otros servidores). Se estima que a fines de 1999 habían al menos siete millones de servidores *web* y probablemente ahora hay más de 14 millones [6,7].

Para comparar la *web* en los distintos países iberoamericanos podemos usar el número de computadores conectados a Internet (en miles). La **tabla 1** muestra algunas estadísticas interesantes, incluyendo el número de miles de computadores por millón de habitantes. En esta tabla cabe destacar el liderazgo inicial de Chile y Costa Rica en Latinoamérica (per cápita), liderados hoy en día por Uruguay, que es el único país con tasas similares a las de España y Portugal. En términos absolutos, España pronto cederá el segundo lugar a México, luego de haber cedido el primer lugar a Brasil durante 1999. Por otra parte, países como Perú, Venezuela y otros menores aún no despegan. En promedio, la región está creciendo 54% más rápido que el resto de Internet en los últimos dos años. Con respecto a páginas *web*, mi estimación personal es que hay alrededor de 50 millones Iberoamérica, lo que sería a lo más un 5% de la *web* mundial.
Por otra parte nos gustaría distinguir servidores físicos de

País	Población (millones)	Internet *hosts* (1998)	Internet *hosts* (2000)	Crecimiento en 6 meses (%)	*Hosts* por millón hab.
Argentina	34.2	20	142	39	4.1
Brasil	159.1	117	446	44	2.8
Chile	14.0	18	40	25	2.9
Colombia	34.6	10	41	32	1.2
Costa Rica	3.1	3	7	103	2.3
España	39.3	169	416	38	10.6
México	89.6	42	405	81	4.5
Perú	23.5	3	9	18	0.4
Portugal	9.9	40	91	54	9.2
Uruguay	3.1	10	25	100	8.1
Venezuela	21.4	4	14	52	0.7
Resto Iberoamérica	72.8	10	19		0.3
Total	**510.5**	**446**	**1655**		**3.2**

Tabla 1. Internet en Iberoamérica

sitios *web* de entes lógicos. Es decir, contar instituciones con servidores *web*, ya que cada institución se puede considerar como una fuente de información distinta. En 1995, el número de sitios era estimado en 30% de los servidores [4] y esa fracción debería haber aumentado si la tasa de crecimiento de nuevas instituciones es mayor que la tasa de crecimiento de nuevos servidores. En el resto de este artículo mencionamos los desafíos inherentes a la búsqueda de información en la *web*, caracterizamos algunos aspectos de la *web*, incluyendo su impacto en Iberoamérica y el castellano, y concluimos con las formas que existen para buscar en ella y cómo funcionan los buscadores de información en la *web*.

2. Desafíos

Buscar información en la *web* implica lidiar con una serie de problemas de distinto tipo. Estos los podemos dividir en intrínsecos a los datos y a los usuarios. Los primeros son:

- **Distribuidos**: dada la estructura de la *web*, los datos están en muchos computadores y plataformas distintas. La topología de la red no está predefinida y el ancho de banda y confiabilidad de las conexiones es muy variable.
- **Volátiles**: los nombres de dominio y páginas aparecen y desaparecen diariamente de la red. Se estima que el 40% de la *web* cambia cada mes. Además el volumen de los datos crece exponencialmente, doblando su tamaño en meses.
- **Dinámicos**: actualmente la gran mayoría de las páginas se generan mediante una consulta a una base de datos y por ende es difícil recuperarlas sin conocer su estructura.
- **Sin estructura**: muchas personas hablan de la *web* como un hipertexto sin ser exactamente cierto. Un hipertexto tiene un modelo conceptual de la estructura y los enlaces de las páginas. Esto difícilmente ocurre en la *web*, y si ocurre es sólo en algunos sitios y de manera distinta. Por eso se habla de datos semi-estructurados.
- **Redundantes**: Una gran cantidad de la *web* esta repetida. El número de *mirrors* (sitios replicados) es de alrededor del 30%. Una cifra similar de páginas han sido parcial o totalmente duplicadas y también hay redundancia semántica (de contenido).
- **Tipos heterogéneos**: hay múltiples tipos medios digitales, de cada medio hay distintos formatos (por ejemplo, HTML o *Word* para texto, o JPG y GIF para imágenes). Además hay diferentes lenguajes y distintos alfabetos, algunos de ellos muy grandes (como *Kanji*).
- **Calidad heterogénea**: la *web* es un nuevo medio de publicación, en muchos casos sin ningún tipo de proceso editorial. Por lo tanto la información de una página puede ser falsa, inválida (es muy antigua), mal escrita, o con muchos errores de diversos tipos. Por ejemplo, en palabras difíciles de escribir la mitad de las ocurrencias pueden estar mal.

Muchos de estos problemas no tienen solución técnica y algunos no debieran ser resueltos (por ejemplo, la diversidad cultural). Además de todo esto, supondremos que una página *web* es lógicamente un documento, lo que no es siempre cierto. Hay documentos que pueden estar en muchas páginas y hay páginas con varios documentos (por ejemplo, resúmenes de los artículos de una revista).

Con respecto a los usuarios tenemos dos problemas básicos: como especificar lo que queremos recuperar (es decir, cual es el lenguaje de consulta) y aunque hayamos especificado exactamente lo que queremos, como manejar respuestas que muchas veces contendrán miles de documentos. Esto implica jerarquizar bien las respuestas. Adicionalmente, algunos documentos pueden ser muy grandes y habría que facilitar el poder examinarlos.

3. Caracterizando la *web*

Estructura y Visibilidad

¿Cuántas referencias tiene una página HTML? (Como se sabe, HTML es un acrónimo de *Hyper Text Markup Language*, el lenguaje usado 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 páginas en el mismo servidor.

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. Si contamos sitios que referencian a sitios, aparece *Yahoo*! en el primer lugar (ver al final una lista de los buscadores y directorios citados en este artículo). 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. Asimismo, 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 800 millones de páginas a comienzos de 1999 [3].

Tamaños y características

¿Cómo es una página *web* promedio? Una página de HTML promedio tiene alrededor de 5 a 7 KB (kilobytes; alrededor 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. En otras palabras, aunque la mayoría de los archivos son pequeños, existe un número no despreciable de archivos grandes, y hasta los 50 KB predomina el volumen de las

imágenes. Desde allí hasta 300 kilobytes son importantes los archivos de audio. Más allá de este límite, 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, PDF, PS y JPG, entre otros.

¿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 mecanografiado, errores que vienen de la conversión de imágenes de documentos a texto, 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 o la referenciamos.

Los Idiomas en la *web*

Existen sólo tres estudios de los distintos idiomas usados en páginas *web*. Uno es de Funredes, una organización no gubernamental establecida en República Dominicana y dirigida por Daniel Pimienta, un francés. Este estudio está hecho en base a frecuencia de palabras en AltaVista y sus últimos datos son de Septiembre de 1998. El segundo estudio pertenece a *Alis Technologies* (*http://babel. alis. com:8080/palmares.html*), una compañía francesa, que hizo un muestreo de 8000 servidores, usando un producto propio que reconoce distintos idiomas. Uno de los objetivos de esta investigación fue validar esta herramienta y data de abril de 1997. En ambos casos, el estudio se centra en el uso del francés. El tercer estudio pertenece a [10] y estima el número de servidores *web* en cada idioma y fue hecho muestreando el 0.1% de las direcciones de Internet en Junio de 1998. La Tabla 2 muestra los resultados más importantes, incluyendo cuántas personas hablan cada idioma. Actualmente, existen páginas *web* en más de 100 idiomas distintos.

De acuerdo a esto, el castellano es la quinta lengua más usada en la *web*, pero debería estar mejor ubicada de acuerdo al número de personas que la hablan (aunque en la **tabla** no aparecen ni el chino ni otros idiomas más hablados cuyos porcentajes son aún menores). Estos datos son aproximados, pues ninguna metodología es exacta y hay muchas páginas multilingües. Otro dato interesante es que no todas las páginas usan ASCII extendido (acentos, etc), y el porcentaje de páginas correctamente escritas es alrededor de 80% en francés y sólo 50% en castellano. De acuerdo al estudio de Funredes, desde 1996 a la fecha la razón francés/castellano ha pasado de 2.4 a 1.1, por lo que este año 2000 el castellano debiera ocupar ya el cuarto lugar.

4. ¿Cómo buscar en la *web*?

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. Los directorios más grandes tienen más de 100 mil categorías jerarquizadas y más de un millón de sitios *web* clasificados. 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, Fast, Inktomi, Northern Light, Lycos* o *Google*, 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. Los problemas principales son la *polisemia*, es decir, palabras que tienen más de un significado, y por lo tanto encontramos páginas que no queremos; y la *sinonimia*,

Idioma	Funredes (%)	Alis Tech. (%)	OCLC	Hablantes (millones)
Inglés	76.4	82.3	71	450
Japonés	4.8	1.6	4	126
Alemán	4.4	4.0	7	118
Francés	2.9	1.5	3	122
Castellano	2.6	1.1	3	266
Italiano	1.5	0.8	1	63
Portugués	0.8	0.7	2	175

Tabla2. Los idiomas en La Red

palabras distintas que tienen el mismo significado y por ende si no usamos la palabra correcta, no encontramos lo que queremos.

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 1998 de *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 él que no mencionan ni *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 de prestigio. 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 encontramos 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 búsquedas por temas, pero también aquí tenemos el problema de la poca amplitud de temas (buscando jaguar sólo se obtienen autos o equipos de fútbol). Búsquedas en *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.

Actualmente no es posible distinguir los buscadores de los directorios, porque los primeros han agregado jararquías y los segundos permiten búsquedas en toda la *web* usando el servicio de algún buscador.

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 (*http://espanol.yahoo.com/*), con páginas específicas de 6 países, entre ellas Chile y España.

Por otra parte, hay otros buscadores especializados. Por ejemplo, en España hay más de 35 de ellos, tales como Ole,

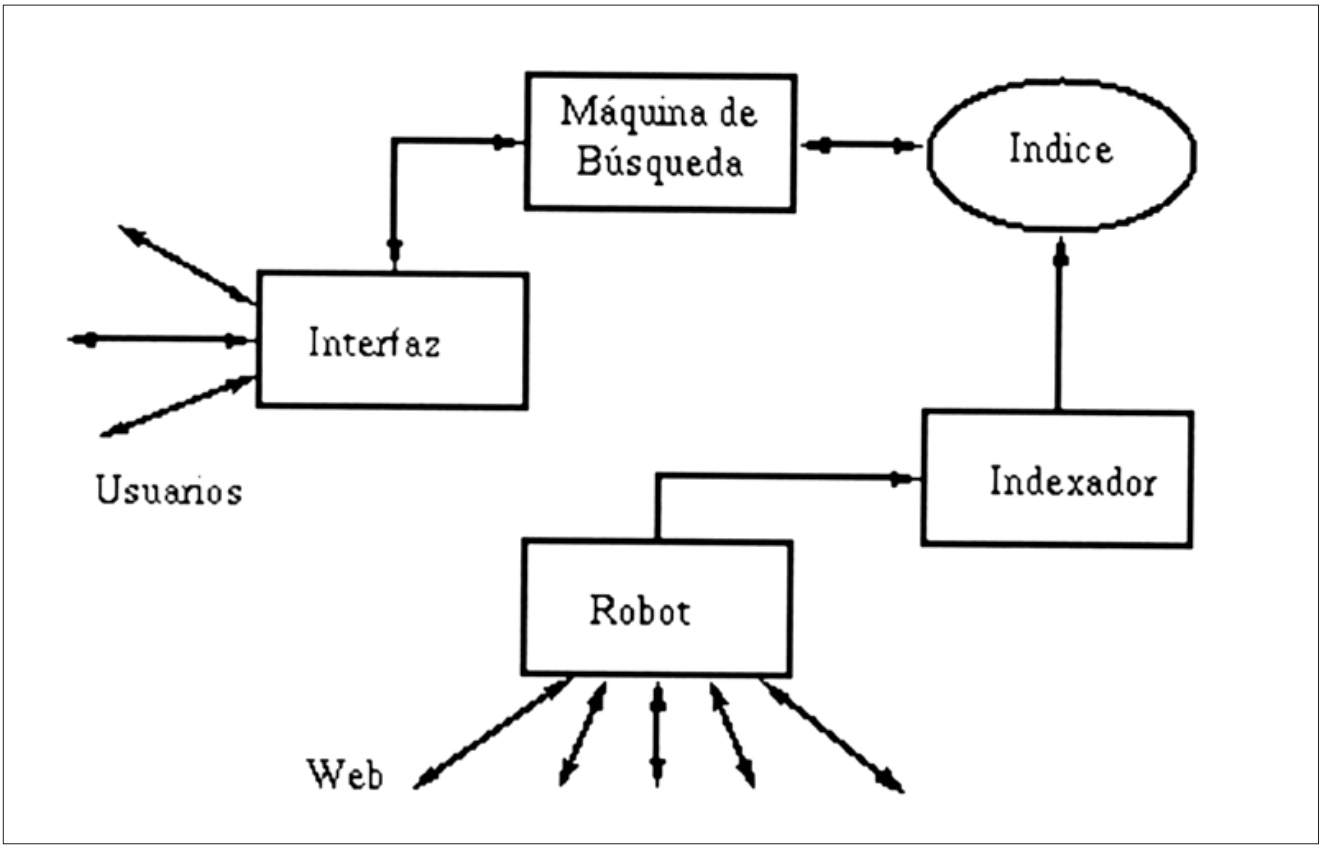

Figura 1. Arquitectura típica de un buscador

Lycos España, BIWE, etc. En Chile un buen directorio es La Brújula mientras que TodoCL es un buen buscador. Este último pertenece a una familia de buscadores Iberoamericanos como TodoBR que permite buscar en toda la *web* brasileña.

5. 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 (**figura 1**) incluye el indexador y el robot. A continuación hablamos de cómo crear un índice de toda la *web*.

Nadie conoce el volumen actual de la *web*. Tratemos de subestimar la cantidad de texto existente en la *web*. Si cada página tiene 5KB y hay como 1.500 millones de páginas, estamos hablando de más de 7.5 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*.

El número de palabras distintas no crece en forma proporcional al texto, sino que crece en forma sublineal (crece como n^x con $0<x<1$). Esto se debe a que el vocabulario es finito y entonces muchas palabras se repiten. Por otra parte, la frecuencia de las palabras sigue una variante de la **Ley de Zipf** que caracteriza la ocurrencia de palabras en el texto. 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).

Usando distintas técnicas, el tamaño de un archivo invertido puede reducirse a un 20% del tamaño del texto. Estos índices se pueden reducir usando particiones lógicas en vez de documentos (por ejemplo, poniendo muchas páginas pequeñas en un mismo grupo).

Usando una búsqueda eficiente 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 uno de los más grandes, registra sobre 300 millones de páginas *web*, y para atender las consultas se usan decenas de servidores Alpha, cada uno con varios procesadores y 8 GB (gigabytes) de memoria RAM (sí, leyó bien, 8GB). Por lo tanto, gran parte del índice y muchas de las respuestas están almacenadas ya en RAM (para poder rápidamente retornar la siguientes 20 páginas de una consulta).

Los otros buscadores con un número similar de páginas son *Fast* e *Inktomi*. Este último recientemente dice haber llegado a los 500 millones de páginas, que significaría una cobertura de alrededor de un tercio de la *web*. Este esquema centralizado tiene un límite si la *web* sigue creciendo como hasta ahora y el final de los buscadores existentes hoy en día podría ocurrir en un futuro cercano.

Resultados recientes demuestran que el número de páginas que están en los buscadores más grandes es pequeño (del 2%

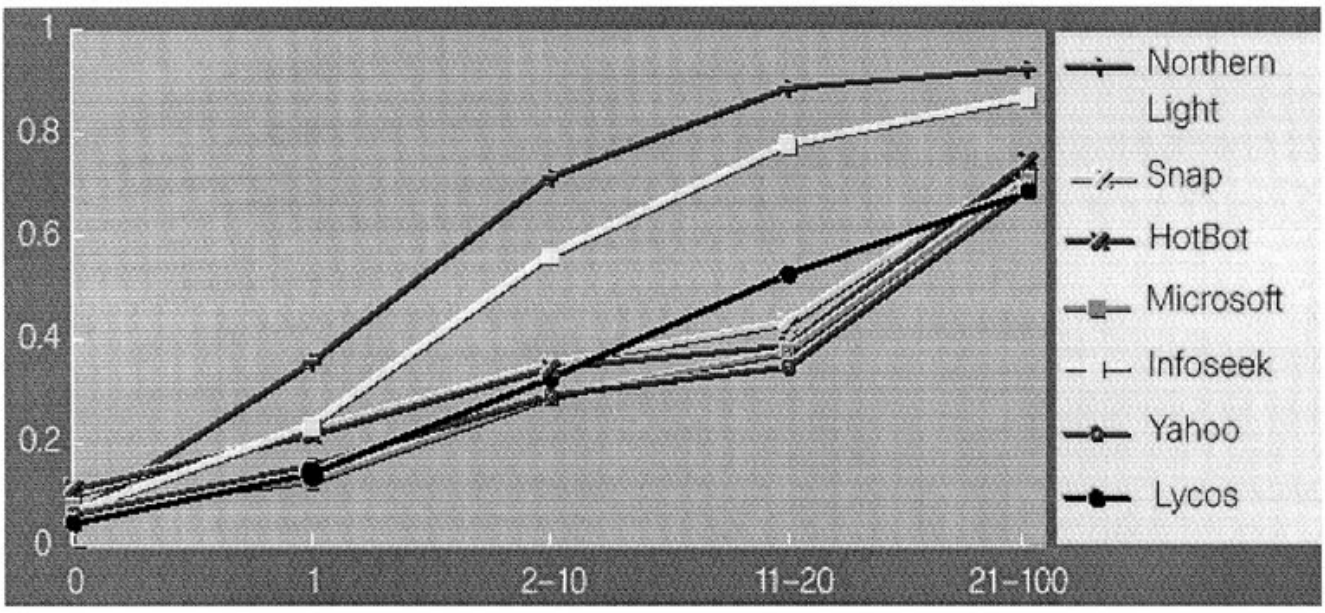

Figura 2. Probabilidades de encontrar páginas

al 5%) y en general se encuentran páginas distintas en cada uno de ellos. Por lo tanto, un buen metabuscador (buscador que busca en muchos buscadores) puede ser muy efectivo si sabe combinar y clasificar bien las respuestas. Otras ideas recientes incluyen agentes de software especializados o metabuscadores en temas específicos, por ejemplo *Search Broker* o *Meta Miner*.

Un problema técnico importante es como jerarquizar las páginas. La mayoría de los buscadores usan la ocurrencias de las palabras que estamos buscando, pero esto muchas veces no funciona. Nuevas técnicas incluyen información de los enlaces, lo que es muy efectivo. Un buscador que usa esta idea es *Google*.

Otro peligro es que los buscadores de Internet estén jerarquizando las respuestas en base a razones económicas y no de contenido. Esto es difícil de demostrar, pero algo igualmente preocupante se muestra en [3] acerca de este tipo de inequidad. Es mucho más probable que una página popular esté en el índice de un buscador como Altavista que una que no lo sea. Una forma de medir la popularidad es contar el número de enlaces dirigidos a una página. La **figura 2** muestra como la probabilidad de encontrar una página en un buscador aumenta dependiendo del número de enlaces que se dirigen a ella (en vez de ser constante).

En este sentido *NorthernLight* es el más equitativo. Así es que ya sabe, pídale a sus amigos que lo apunten y aparecerá en más buscadores.

6. Conclusión

La *web* es un gran repositorio de datos y un nuevo medio de publicación al alcance de más de 100 millones de personas. El hacer uso eficiente y adecuado de estos datos depende de nosotros y de las herramientas que existen y que han sido descritas en este artículo. El futuro dirá si es posible adaptar estas herramientas al crecimiento explosivo de la *web* y que además la *web* misma no colapse debido a la congestión en las redes y servidores *web*. Para mayor información en este tema, ver el capítulo 13 de *Modern Information Retrieval* [2].

En al ámbito iberoamericano, seguirán apareciendo más buscadores, pues una solución al problema es verticalizar la búsqueda. Una forma de hacer esto es geográficamente por países o también culturalmente por idiomas.

7. 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, Wokingham, Inglaterra, Marzo 1999. (*http://sunsite.dcc.uchile.cl/irbook*)
[3] S. Lawrence y L. Giles, *Accessibility of Information on the web*, Nature, Julio 1999.
[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.searchenginesshowdown.com/*)
[10] OCLC, *Study of web Characteristics*, 1998. (*http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm*)
[11] *Scientific American*, Número especial dedicado a Internet, Marzo de 1997. (*http://www.sciam.com/0397issue/0397currentissue.html*)
[12] Danny Sullivan, *Search Engine Watch: Main Page*, 1997. (*http://www.searchenginewatch.com/*)

Buscadores y directorios citados en este artículo

Altavista: http://www.altavista.com/
Fast: http://www.alltheweb.com/
Funredes: http://funredes.org/
Google: http://www.google.com/
Inktomi: http://www.inktomi.com/
La Brújula: *http://www.brujula.cl/*
Lycos: http://www.lycos.com/
Meta Miner: http://www.miner.com/
Northern Light: http://www.nlsearch.com/
Search Broker: http://debussy.cs.arizona.edu/sb/
TodoBR: *http://www.todobr.com.br/*
TodoCL: *http://www.todocl.cl/*
Yahoo: http://www.yahoo.com/