Características de la Web Chilena 2006

Ricardo Baeza-Yates
Yahoo! Research
Centro de Investigación
de la Web

Carlos Castillo
Yahoo! Research

Eduardo Graells
Centro de Investigación
de la Web

Marzo de 2007

Resumen Ejecutivo

En Agosto de 2006 se llevó a cabo una recolección masiva de páginas de la Web de Chile utilizando el sistema WIRE, desarrollado en el CIW. Del análisis de estos datos destacan las siguientes observaciones:

- La Web chilena está compuesta por más de **170.000 sitios**, y estos sitios contienen más de **7 millones de páginas**. Muchas de sus características son muy similares a las de la Web global en general.
- Un **14 % de los sitios están conectados entre sí** a través de enlaces y tienen **el 53,3 % de las páginas**. Por otro lado, un 49,5 % de los sitios está completamente desconectado en términos de enlaces, pero representan sólo el 14 % de las páginas.
- Un sitio promedio tiene **43 páginas**, contenidas en **0,304 MiB**, con **1,56 referencias desde otros sitios**.
- Un dominio promedio tiene **1,08 sitios** y **46,61 páginas**, contenidas en **0,328 MiB**.
- Cerca de **1/4 de las páginas chilenas fue creada o actualizada en el último año**, lo que implica un alto grado de crecimiento y dinamismo.
- Alrededor del 80 % de las páginas de Chile está en español y cerca de un 17 % en inglés. Otros idiomas tienen una presencia muy leve.
- Los sustantivos que más aparecen en la Web chilena son: Chile, producto, usuarios, servicio y mensaje. También aparecen Santiago, Web, blog, región e información.
- Los países más referenciados desde Chile son Argentina, España, Alemania, Reino Unido y México, y en general el número de referencias a países extranjeros está relacionado con el volumen de intercambio comercial.
- Los sitios que reciben más enlaces son `sii.cl`, `uchile.cl`, `mineduc.cl`, `meteochile.cl` y `bcentral.cl`.

- Los proveedores de *hosting* con mayor número de sitios son IFX Networks, VirtuaByte, T–Chile, Telefónica Internet Empresas, DattaWeb y PuntoWeb.

Respecto a la calidad de las páginas y sitios:

- De todos los sitios, el **20 % más grande de ellos contiene el 99 % de la información en la Web chilena**, medida en el número de *bytes* contenidos en sus páginas.
- Cerca de un **21 % de los sitios de Chile no son fáciles de encontrar** ya que están hechos con tecnologías no visibles para los motores de búsqueda, como Flash y Javascript.
- Unas pocas páginas acaparan la mayoría de los enlaces. De hecho, sólo un 3 % de las páginas tienen algún valor de contenido en términos de estar referenciadas desde otros sitios. Sin embargo, estas páginas están repartidas en el 35 % de los sitios Web.
- Cerca de un **5 % de los enlaces ya no existen**.

Respecto a las tecnologías Web:

- De los servidores que entregan información, el servidor Web más utilizado es *Apache* con 66,7 %, seguido con un 32,8 % por *Microsoft Internet Information Server*.
- De los servidores que entregan información, el sistema operativo más utilizado es *Unix* con 48,5 %, seguido por *Microsoft Windows* con 38,5 %. Además, *Linux* es utilizado en un 12 % de los servidores.
- El generador de páginas dinámicas más usado es PHP con un 75 % de participación en el mercado.
- El formato de documentos más usado es PDF con un 53 % de participación, seguido por XML con un 21 %.
- Aproximadamente hay una disponibilidad del doble de archivos con paquetes de software para Linux que para Windows en la Web chilena.

Índice

1. Introducción

En esta sección presentamos las características de la Web y de la muestra estudiada, así como la metodología para recolectar documentos.

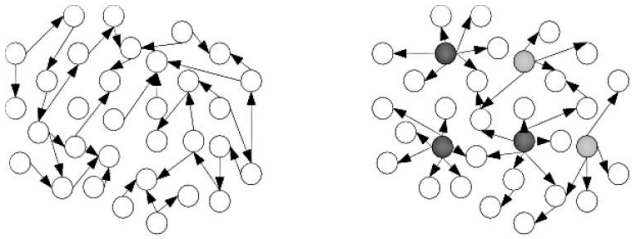
1.1. ¿Cómo es la Web?

La Web es más que un simple conjunto de documentos en distintos servidores, ya que existen relaciones de información entre los documentos mediante los enlaces que establecen entre ellos. Esto presenta muchas ventajas, tanto para los usuarios, a la hora de buscar información, como para los programas que recorren la Web, a la hora de buscar contenido para recolectar (probablemente para un motor de búsqueda). Debido a esto se plantea que la Web sigue un modelo de grafo dirigido, en el que cada página es un nodo y cada arco representa un enlace entre dos páginas.

En general las páginas enlazan a páginas similares [19], de modo que es posible reconocer páginas mejores que las demás, es decir, páginas que reciben un número mayor de referencias que lo normal. Como se explicó anteriormente en [9], la web tiene una estructura que se puede clasificar como *red libre de escala*. Dichas redes, al contrario de las redes aleatorias, se caracterizan por una distribución dispareja de enlaces y porque dicha distribución sigue una ley de potencias (*power-law*)[1]:

$$P_r\left(\Gamma\left(p\right)=k\right)\alpha k^{-\theta}$$

Los nodos altamente enlazados actúan como centros que conectan muchos de los otros nodos a la red, como se ilustra en la Figura 1.



(a) Red aleatoria (b) Red libre de escala

Figura 1: Ejemplos ilustrativos de una red aleatoria y una red libre de escala. Cada grafo tiene 32 nodos y 32 enlaces.

Esto quiere decir que la distribución de los enlaces es muy sesgada: unas pocas páginas reciben muchos enlaces mientras que la mayoría recibe muy pocos o incluso ninguno. En este estudio se muestra que dicha distribución se puede aplicar a muchos aspectos de la Web, de los cuales se dice que "siguen una ley de Zipf", en honor a Kingsley Zipf que en 1932 enunció la distribución que modela la frecuencia de aparición de las palabras en los textos [43]. De acuerdo a este modelo la probabilidad de encontrar un elemento con un cierto tamaño x es proporcional a x^{α}, con $\alpha > 1$. Cuando estas distribuciones se representan en un gráfico con escala logarítmica se obtiene una línea recta, tal como se observa en muchos de los gráficos de este estudio.

[1] Respecto a su estudio se recomiendan el trabajo de Barabási [13] debido a su claridad.

1.2. Estudiando la Web de un país

Las redes libres de escala son auto-similares: una pequeña muestra tiene características de la red completa (es decir, las características trascienden la escala con que se mire la red). Se muestra en este estudio que éste es el caso de la Web chilena, que presenta características muy similares a la red mundial y a las redes de otros países, a pesar de contener menos de 2/3000 de las páginas recolectables[2] en la Web global, estimadas el 2005 en $11x10^9$ páginas [27].

Una Web nacional se puede definir como el conjunto de páginas relacionadas con un país. Técnicamente es difícil distinguir si una página está asociada al país en estudio; debido a ello, para el caso chileno se utiliza la heurística de asociar a Chile todos los sitios Web que están hospedados en direcciones IP asignadas a Chile. Esto incluye los dominios chilenos (`.cl`) y otros dominios genéricos y extranjeros. Conocemos todos los dominios `.cl` gracias al apoyo de NIC Chile (`http://www.nic.cl`), que entregó la lista completa de dominios para fines de investigación.

La Web chilena ha sido objetivo constante de estudio, tanto de sus características en los años 2000 [3], 2001–2002 [12] y 2004 [5], como de sus dinámicas [11]. Asimismo, existen recientes estudios sobre las Webs de otros dominios nacionales:

- África (9 países) [16]
- Argentina (sólo universidades) [41]
- Austria [36]
- Brasil [32, 42]
- China [30]
- España [9]
- Grecia [21]
- Hungría [14]
- Corea del Sur [10]
- Perú [40]
- Portugal [24]
- Reino Unido, Nueva Zelanda y Australia (sólo universidades) [39]
- Tailandia [37]

[2]La Web pública, e indizable, es sólo una parte de la Web total. La Web oculta, de acceso restringido o privado, probablemente es mucho más grande.

1.3. Recolección de páginas

La colección fue obtenida el mes de Agosto de 2006, utilizando el programa para recolectar páginas web WIRE [7]. Se utilizó un computador con dos procesadores Intel Pentium IV de 3 GHz, 1 GiB[3] de RAM bajo sistema operativo Gentoo Linux.

El recolector comienza descargando un conjunto de direcciones iniciales (*seeds*), que corresponden a la lista de dominios que se posee. De las páginas descargadas se extraen nuevos enlaces, de los cuales se discriminan los que no apuntan a páginas o documentos en dominios chilenos o a direcciones IP que no estén asignadas a proveedores chilenos. En total se descargaron más de 7 millones de páginas web (más del doble que las descargadas el año 2004 [5]). La colección utiliza 50 GiB de disco, de los cuales 48 GiB corresponden al texto de los documentos y 2 GiB a metadatos de las páginas.

El Cuadro 1 resume las características principales de la colección.

Páginas Web	7.403.840
Texto en total	48,56 GiB
Texto promedio por página	7,04 KiB
Sitios Web	171.213
Páginas promedio por sitio	43,24
Texto promedio por sitio	304,59 KiB
Dominios	158.853
Sitios promedio por dominio	1,08
Páginas promedio por sitio	46,61
Texto promedio por dominio	328,29 KiB

Cuadro 1: Cuadro Resumen de la Colecta.

1.4. Dificultades en la caracterización de la Web

La Web es una colección descentralizada, en la cual distintos autores pueden contribuir contenido por su cuenta sin una instancia de control que decida qué se publica y qué no. Esto es la principal ventaja de la Web desde el punto de vista de los usuarios, pero también es la principal causa de dificultades tanto para buscar información como para caracterizar colecciones de páginas.

Las siguientes anomalías constituyen violaciones de estándares o situaciones especiales que dificultan la caracterización de las páginas:

Parámetros en la URL y URL Rewriting : existen páginas que tienen direcciones más largas de lo que realmente deberían ser. Esto se debe a que entregan sus parámetros en la dirección de la página como si fuera la ruta de acceso a ella, lo que contradice el estándar de URLs [15], puesto que los parámetros de invocación de programas deberían aparecer en la URL después de un signo "?", por ejemplo:

- Incorrecto: `http://sitio/directorio/buscar/palabra/X/maximo/10/`

[3]Usamos "GiB", "MiB", etc. para referirnos a potencias en base 2, mientras que "GB" y "MB" se refieren a potencias en base 10.

- Correcto: `http://sitio/directorio/buscar?palabra=X$&$maximo=10`

Esta técnica es conocida como *URL Rewriting* y su uso se ha extendido con la aparición de sistemas de administración de contenido (CMS, Content Management System). Entre sus consecuencias se encuentran: *1)* no se puede distinguir si la página es estática o dinámica, y *2)* se recorren varias páginas que tienen semánticamente el mismo contenido, ya que por lo general estas direcciones admiten varios parámetros diferentes para entregar una misma página (el identificador, el título, la sección dentro del sitio, la fecha, etc.). Así se encuentran sitios que tienen un tamaño mucho más grande del que tienen realmente, con muchas más páginas que el promedio.

Réplicas de contenido : Constituye una práctica habitual en la Web el tener varias copias distribuídas geográficamente de los mismos documentos. Normalmente lo que se replica son colecciones completas de gran volumen, y se hace por motivos de eficiencia. Las colecciones más frecuentemente replicadas en la Web son [18]: el sitio de software Tucows, el proyecto de documentación de Linux (LDP), la documentación del servidor web Apache y la documentación del lenguaje de programación Java. La información replicada se estima entre un 20 % y un 40 % del total en la Web.

Las consecuencias de estas réplicas, sitios con una gran cantidad de texto, son pequeñas considerando la realidad de la Web chilena, donde los documentos que son réplicas de otros son 232.705, un 3,14 % del total de las páginas. Una inspección manual de la colección nos lleva a observar mucho contenido repetido que no se detecta como duplicado debido a que el contenido de las páginas también incluye el diseño, y éste por lo general varía a pesar de presentar el mismo contenido. Existen muchos sitios que duplican el contenido entre ellos de forma deliberada y no por las razones de eficiencia o documentación.

Spam en general : El Spam en la Web se refiere a acciones orientadas a engañar a los sistemas de búsqueda en la Web y a dar algunas páginas una posición más alta de la que merecen en el resultado de una búsqueda en un motor de búsqueda [28]. Estas acciones incluyen cambios en el texto, en los metadatos o en los enlaces de las páginas si es que el visitante es un robot recolector.[4]

1.5. Organización de este informe

Los distintos niveles de análisis posibles para la Web son los siguientes: el más pequeño es el de palabras o bloques de texto o imágenes, pasando por páginas, sub-sitios (unidades coherentes de múltiples páginas), sitios, dominios; hasta llegar a la Web de un país y la Web global. Así está estructurado este informe, presentamos nuestras observaciones de la Web de Chile a varios niveles: a nivel de páginas y documentos en la Sección 2, a nivel de sitios en la Sección 3 y a nivel de dominios en la Sección 4. La Sección 5 presenta nuestras conclusiones, y el Glosario incluye definiciones de algunos términos que se utilizan en este documento

[4]Reconocer fehacientemente cuáles páginas son spam o no es un área activa de investigación, y se estima que hasta un 8 % de lo que indizan las máquinas de búsqueda en la Web es spam, siendo uno de los signos destacados del spam la presencia de regularidades en la distribución de ciertas variables [22].

2. Características de las Páginas

En esta sección se presenta el análisis de las páginas individualmente, sin cosiderar su agrupación en sitios o dominios. Primero enseñamos el número de páginas descargadas correctamente. Luego analizamos los metadatos, como la URL, el título o el tamaño, más tarde el contenido de los documentos y finalmente los enlaces entre ellos.

2.1. Páginas descargadas versus enlaces inválidos

El recolector de páginas funciona extrayendo direcciones de las páginas que han sido descargadas, y es frecuente que entre estas direcciones aparezcan páginas que ya no existen o que simplemente se escribieron mal. Cada vez que el recolector se contacta con un servidor Web, éste retorna un código de estado que indica si la página existe o no, o si existe un motivo por el cual no se puede entregar el documento pedido. La Figura 2 muestra la distribución de páginas de acuerdo a estos códigos de estado. Existe una gran cantidad de códigos de estado que se han agrupado de la siguiente manera:

- `OK`: incluye todos los requerimientos exitosos: `OK (200)` y `PARTIAL CONTENT (206)`.
- `NOT FOUND`: el servidor no encuentra el documento pedido: `NOT FOUND (404)`.
- `MOVED`: incluye todos los requerimientos en los cuales el servidor redirige al recolector a una otra página: `MOVED (301)`, `FOUND (302)` y `TEMPORARY REDIRECT (307)`.
- `SERVER ERROR`: incluye todas las fallas en el lado del servidor: `INTERNAL SERVER ERROR (500)`, `BAD GATEWAY (502)`, `UNAVAILABLE (503)`, y `NO CONTENT (204)`.
- `FORBIDDEN`: incluye todos los requerimientos que no son permitidos, principalmente por tratarse de páginas protegidas con clave: `UNAUTHORIZED (401)`, `FORBIDDEN (403)` y `NOT ACCEPTABLE (406)`.

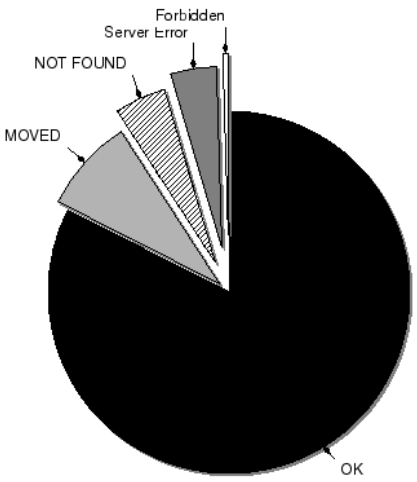


Figura 2: Distribución del código de estado HTTP

En nuestros experimentos, usualmente obtenemos entre 75 % y 85 % de transferencias exitosas. En este caso se está dentro del promedio con un 82,5 %, disminuyendo 4 puntos porcentuales respecto a nuestro último estudio [5]. También disminuyó cerca de 2 puntos la proporción de los enlaces rotos, ahora en un 4,6 %. La disminución de los enlaces rotos puede significar que existe mayor conciencia respecto a verificar la correctitud de los enlaces en un sitio.

2.2. URLs

La dirección de una página Web es comúnmente expresada mediante una URL (*Uniform Resource Locator*) [15]. Las URLs tienen un doble propósito, por una parte identifican un recurso en la Web de manera única y por otra indican cómo es posible acceder a dicho recurso en el servidor.

Las URLs más usadas en la Web son las que corresponden al protocolo de transferencia de hipertextos (HTTP). Estas URLs tienen normalmente la siguiente forma:

`http://sitio/directorio/archivo`

Por ejemplo, `http://www.cwr.cl/projects/WIRE/index.html` indica que el sitio a contactar es `www.cwr.cl`, que el archivo que se necesita se encuentra en el directorio `/projects/WIRE/` y que se llama `index.html`.

En esta sección analizamos algunas de las características observadas de las URLs de la Web de Chile.

2.2.1. Longitud de las URLs

La longitud promedio de una URL, incluyendo la especificación del protocolo `http://`, nombre de servidor, ruta y parámetros, es de 71 caracteres. Este promedio es mayor al de otros países: 67 para España [9] y 62 para Portugal [24], y 50 caracteres para la Web global [38]. Sin duda influyen en esto las nuevas aplicaciones que se están desarrollando en la Web que requieren un gran número de parámetros; a la vez los parámetros tienden a expresarse como texto, como se mencionó en la Sección 1.

El 60 % de las URLs tienen entre 40 y 80 caracteres. Los largos se distribuyen de acuerdo con la Figura 3, que tiene una distribución *log-normal* con parámetros estimados $\theta = 9$ (posición), $m = 53,81$ (escala) y $\sigma = 0,48$ (forma). Observamos que las URLs más largas, a pesar de pertenecer a sitios distintos, comparten un parámetro que por simple inspección parece ser un identificador de sesión o de servidor. Las URLs que siguen corresponden en su mayoría a URLs con elementos repetidos. Las URLs anómalas tienen formas similares a las siguientes (han sido recortadas por motivos de legibilidad):

- `http://www.visitamway.cl/search.php?d=visitamway.cl&term=Merchant+Accounts&nterms=TWVyY2hhbnQgQWNjb3VudHM=&cachekey=03u3hs9yoajl1K3TDEhRDG[...]n1nc9JLmxxcis21BE`
- `http://www.psicodocencia.cl/blog/node/taxonomy/taxonomy/node/taxonomy/[...]/node/taxonomy/taxonomy/node/taxonomy?from=60`
- `http://www.upadiseno.cl/mundo_academico/escuela_de_diseno/fileadmin/mundo_academico/escuela_de_diseno/[...]`
- `http://www.webs.cl/info/122/caracteristicas-de-un-partido-politico/info/122/caracteristicas-de-un-partido-politico/info/[...]`

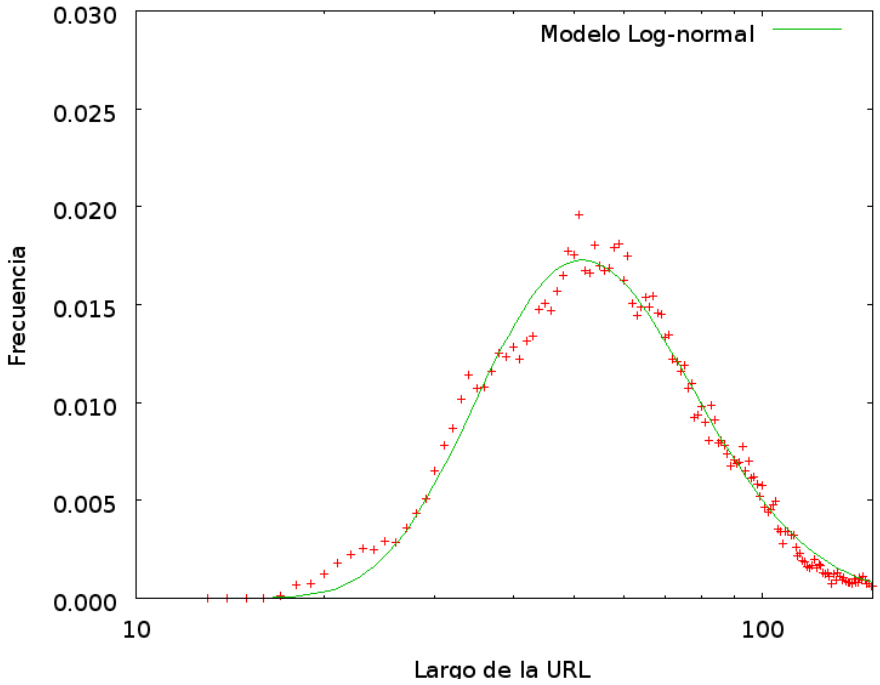


Figura 3: Distribución del largo de las URLs.

2.2.2. Profundidad según URL

Estudiamos la profundidad de una página dentro de un sitio Web. Esta profundidad puede definirse de dos formas:

Profundidad Lógica La página inicial de un sitio está a profundidad 1; todas las páginas alcanzables directamente desde ella, a profundidad 2; y así sucesivamente. La profundidad lógica mide el número de "clicks" necesarios desde la portada de un sitio hasta la página requerida.

Profundidad Física La página inicial de un sitio está a profundidad 1, las páginas de la forma `http://sitio/pag.html` o `http://sitio/dir/` están a profundidad 2, y así sucesivamente. La profundidad física mide la organización en archivos y directorios de un sitio Web.

En este estudio analizamos la profundidad física de las páginas, que es directamente extraíble de las URLs. La distribución de esta variable se muestra en la Figura 4, en la cual hemos separado páginas estáticas y dinámicas. Se observa que el máximo de la distribución se encuentra en los niveles 4 y 5.

Es necesario mencionar que en la configuración del recolector se limitó la profundidad máxima a 15 niveles. Probablemente si se hubiese aumentado a 30 se encontrarían anomalías en la distribución en niveles superiores [9].

2.3. Edad de las páginas

Para determinar la edad de las páginas, observamos la fecha de última modificación entregada por los servidores para cada una de ellas. En algunos casos tal fecha es errónea: corresponde a una fecha del futuro o una fecha muy antigua previa a la invención de la Web (como *1 de Enero de*

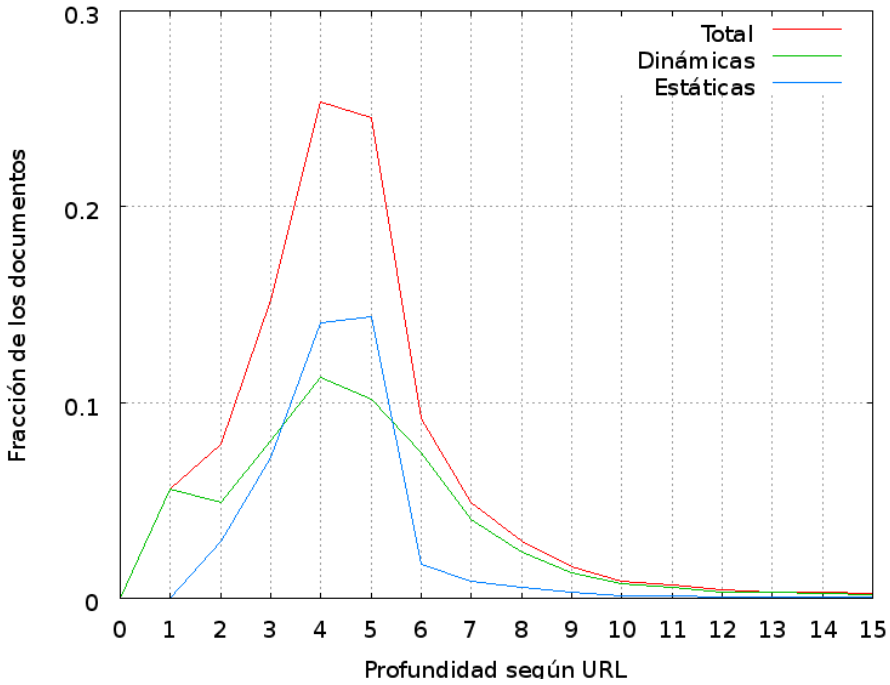


Figura 4: Distribución de páginas a diferentes profundidades, según la URL.

1970). Esto se debe a servidores que no tienen sus relojes sincronizados con la hora actual del país o que simplemente no han sido configurados.

La distribución de las edades de las páginas en términos de meses y años se muestra en la Figura 5.

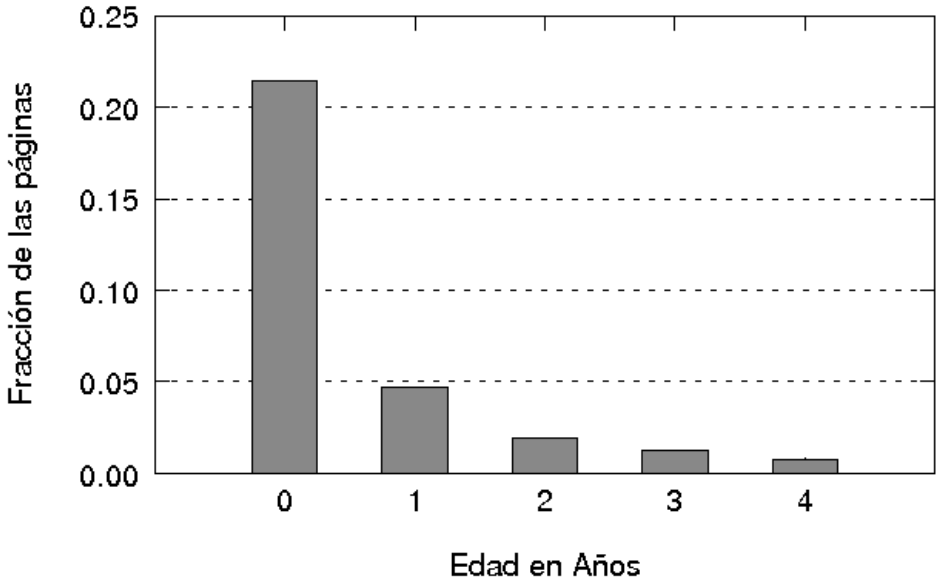
Entre los años 2005 y 2006 cerca de un 25 % de las páginas se creó o se actualizó, lo que indica que la Web chilena está creciendo a una tasa alta, aunque relativamente menor a la tasa de crecimiento del año 2004 (que fue de 25 % sólo para el 2004 [5]).

2.4. Títulos de las páginas

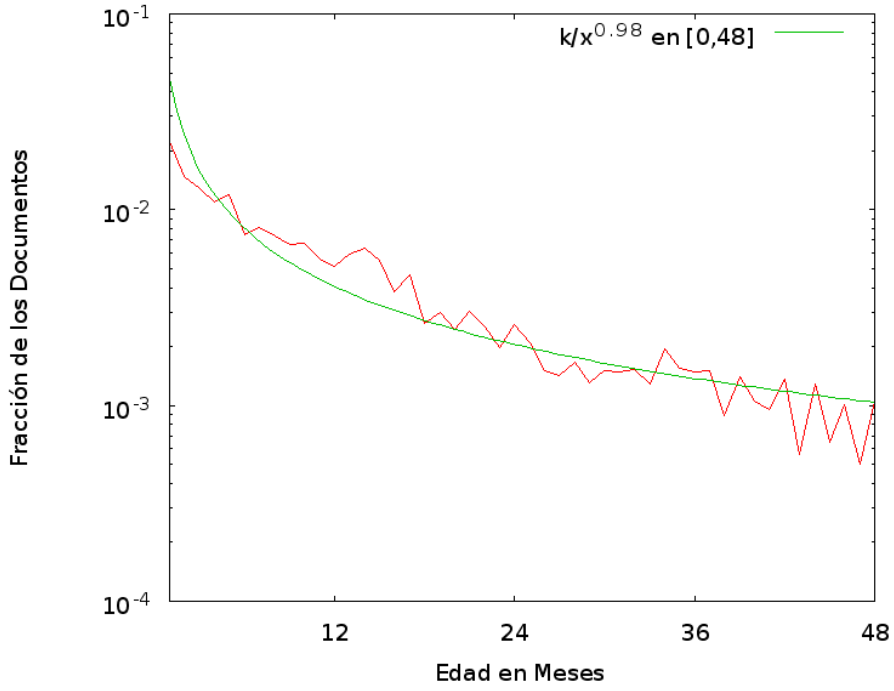
Examinamos el título de las páginas y encontramos que más de un 12,19 % de ellas no tiene título, que un 2,47 % tiene algún título por omisión como "*Untitled Document*", "*Documento sin título*" o "*Página nueva 1*" y que un 85,34 % tiene un título acorde al sitio (51,12 % compartido y 33,22 % único), como se aprecia en la Figura 6.

Lo más común es que una página tenga como título el nombre del sitio al que pertenece, nombre que generalmente no describe el contenido del documento. En otros países este fenómeno es bastance crítico, como en la Web de España donde las mayores frecuencias de largo de título se encuentran entre 5 y 10 caracteres [9]. En la Web chilena, sin considerar a las páginas sin título, la distribución de los largos de ellos no es tan sesgada, como se ve en la Figura 7.

En cada sitio se tienen aproximadamente 6 títulos distintos, una cifra que es buena en comparación con España (4 títulos por sitio [9]) y Portugal (5 títulos por sitio [24]). Eso no quiere decir que no existan anomalías: por inspección manual se aprecia que muchos títulos, en especial los compartidos, no sólo incluyen el nombre del sitio sino que también una descripción de éste. Si bien esto no está mal, no se está utilizando correctamente el atributo de título como un ítem descriptivo conciso del documento (para una descripción más elaborada existe el metadato *des-*



(a) Distribución por años de antigüedad.



(b) Distribución por meses de antigüedad.

Figura 5: Distribuciones de la edad de las páginas web.

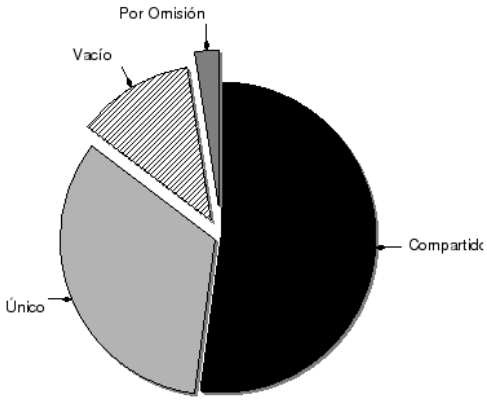


Figura 6: Distribución del título de las páginas: compartido (es usado por más de una página en el mismo sitio), único (es usado exclusivamente por esa página dentro del sitio).

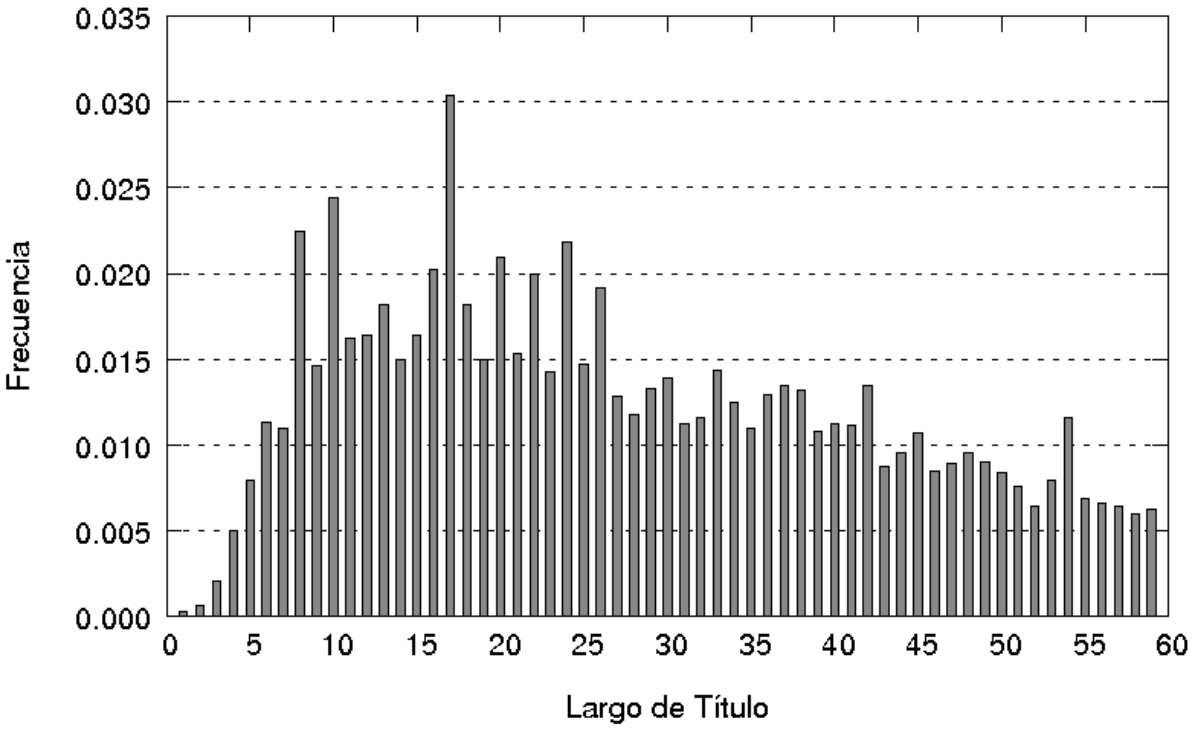
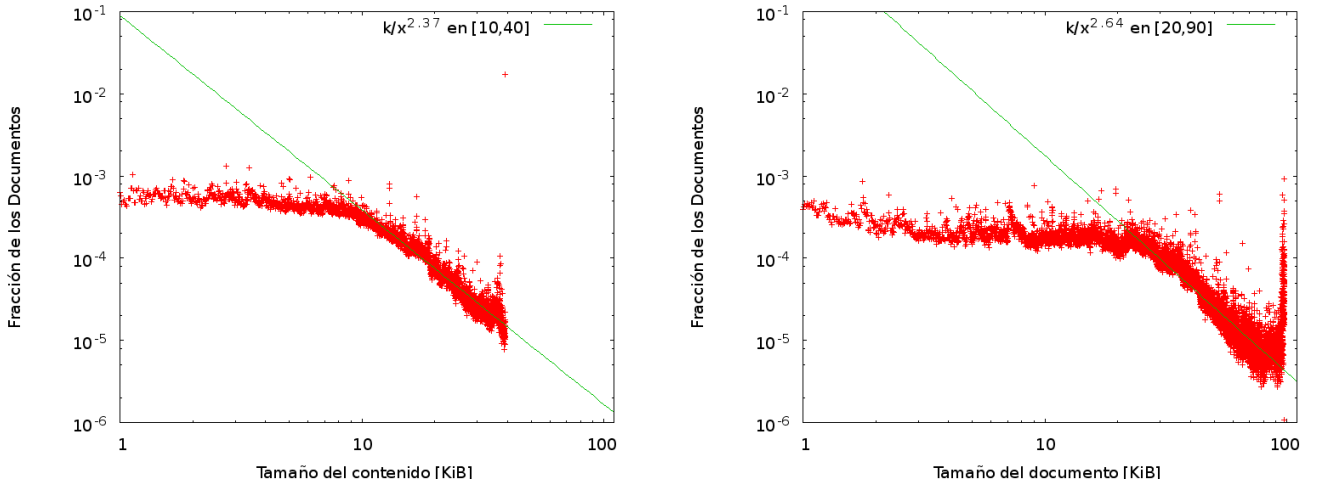


Figura 7: Distribución de los largos de los títulos de las páginas.

cription). Esto incide negativamente en la usabilidad del sitio, porque a quienes visitan el sitio no se les entrega información contextual de los documentos para ser guardados en sus marcadores (*bookmarks*, o enlaces favoritos). Por el lado de los buscadores, si bien las páginas son indizadas correctamente, lo que se muestra al usuario en los resultados de búsqueda es el título de la página y un extracto del documento. Otra anomalía observada es que muchos sitios, en particular aquellos de contenido manejado por los usuarios, contienen caracteres sin relevancia ni significado en los títulos, entorpeciendo su lectura (por ejemplo: "*:::::::: ¡Mi Página ¿:::::::::*").

2.5. Texto en las páginas

De cada página descargada se almacenaron sólo los primeros 100 KiB, lo que es suficiente para la mayoría de ellas. Representamos gráficamente el contenido de las páginas en la Figura 8. Primero se muestra solamente el contenido del documento, y luego el texto completo (contenido más etiquetas HTML y código).



(a) Distribución del contenido (texto sin HTML). (b) Distribución del texto completo (texto con HTML).

Figura 8: Distribuciones del tamaño de las páginas de los sitios web.

El tamaño de los documentos se ajusta a una ley de Zipf con parámetro 2,87, similar al valor obtenido para el año 2004, 2,76 [5]. Respecto a otros países, la mayor similitud se da con Corea del Sur, que posee un valor de parámetro de 2,84 [10].

En la Figura 9 se observan las páginas con menos de 100 bytes de texto. Se observa que entre 0 y 10 bytes hay una similitud con una distribución normal, mientras que el resto del rango se puede modelar con una recta que represente una fracción de las páginas de 3/10000.

Al inspeccionar manualmente estas páginas se observa que varias de ellas corresponden a sitios construídos con elementos gráficos o programas insertados en las páginas (como Adobe Flash o Java). También hay muchos documentos que simplemente no tienen contenido, es decir, su texto es "*< html >< body >< /body >< /html >*".

2.6. Idioma

WIRE incluye un sistema de detección de idioma basado en *stopwords*, es decir, palabras que carecen de significado por sí mismas, también llamadas palabras funcionales. La heurística cuenta

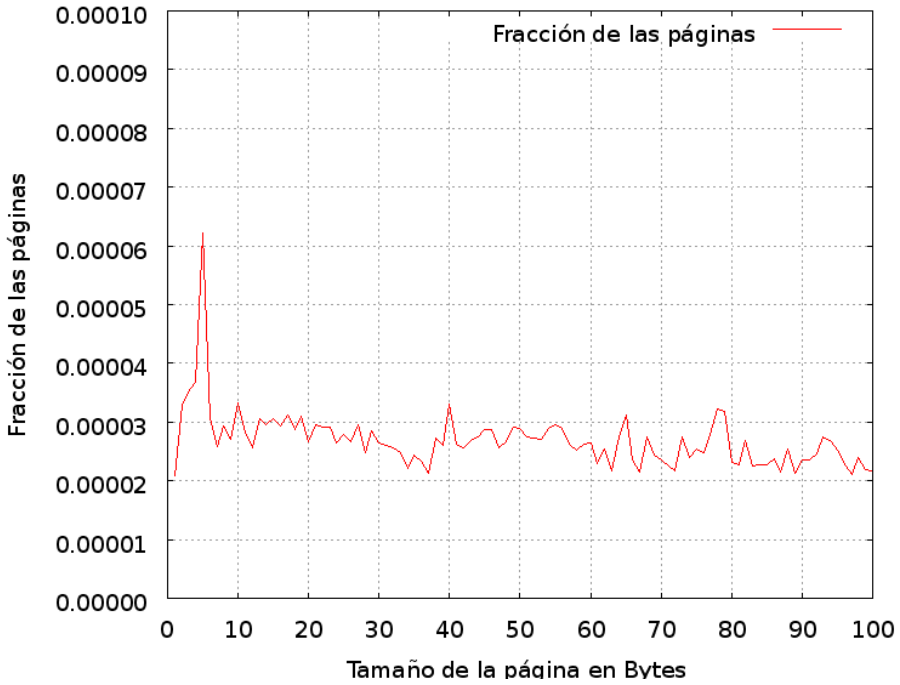


Figura 9: Distribución del tamaño de las páginas, páginas de menos de 100 bytes de texto.

el número de stopwords dentro del documento para cada idioma del que se tienen stopwords y en base a ellas determina el lenguaje correspondiente. En base a esto se obtuvo la distribución de idiomas de la Figura 10.

La proporción de páginas escritas en el idioma oficial del país bajó para Chile en comparación con el estudio del año 2000, de 90 % a 80,21 %; sin embargo, es mayor a la proporción de 62 % presente en España [9], que considera Castellano y Catalán, al 75 % de páginas Brasileñas en Portugués [42], y al 35 % de Tailandés en Tailandia [37]. [5]

2.7. Vocabulario

La definición de palabra que usamos es cualquier secuencia alfanumérica de más de un caracter de largo, incluyendo los caracteres acentuados del idioma español. Esto incluye la conversión a caracteres de entidades HTML (como "`aacute;`" a "`á`"). Analizamos 2,6 GB de texto extraído de páginas de la colección. En la Tabla 2 se indican las diez palabras más frecuentes para los dos idiomas con mayor presencia en la Web chilena.

Español	de	la	y	en	el	que	los	del	por	para
Inglés	the	of	and	to	in	for	by	is	this	on

Cuadro 2: Palabras más frecuentes para los dos idiomas principales.

Naturalmente las palabras más frecuentes son stopwords. En la Figura 11 se muestra la distribución de la frecuencia de las palabras presentes en la colección, obteniendo una ley de Zipf

[5] Este fenómeno, en el cual el Inglés llega a superar al idioma oficial del país en estudio, es tratado con mayor profundidad en [8].

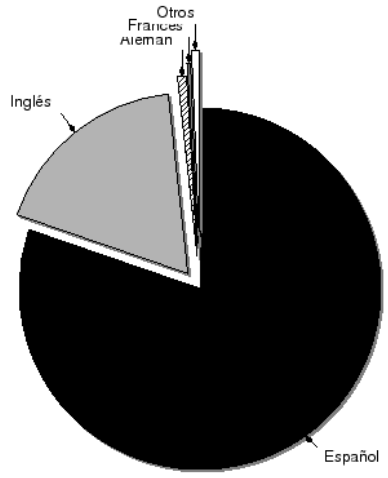


Figura 10: Distribución del idioma en el que están escritas las páginas.

con parámetro 0,84 para el español. En la Web de España el Castellano presenta prácticamente el mismo parámetro [9].

En la Tabla 3 se incluye una Nube de Etiquetas (*TagCloud*, con los sustantivos más frecuentes en la Web chilena. En la Nube de Etiquetas las palabras con mayor presencia tienen un tamaño más grande. Usualmente el término más presente en la Web de un país es el país mismo, y en el caso de Chile se cumple esa regla, al igual que en el caso de Perú (en [40] también se aprecia una nube de etiquetas), España [9] y Brasil [42]. Se observa que, dejando de lado los nombres de ciudades y fechas, en general los términos corresponden a servicios o a tecnologías recientes: en su mayoría son palabras muy frecuentes en foros, blogs, y otros administradores de contenido. Respecto al idioma inglés no se encontró una familia de palabras que perteneciera a un ámbito en particular.

chile producto usuario todos servicio mensaje empresa comentario web santiago blog hora foro septiembre sitio tema noticias región proyecto información precio nombre inicio trabajo universidad agosto tiempo vida cuenta hoy compra mundo lecturas venta desarrollo grupo sistema internet forma derechos bien uso contacto personas centro software julio video fecha mar casa lugar seguridad fotos e-mail página estado miembro programa salud medio copyright datos argentina año respuesta ofertas comunidad historia caso condiciones juegos mayo junio total personal amigo rss clave social red equipo calidad ayuda hotel poder eventos semana internacional actividades libre gobierno fax artículos palabras control momento guía ciudad ley abril día sociedad alumnos click publicaciones responsable central favor accesorios dvd cultura viernes resultados director libro correo post digital gente búsqueda detalles marzo server online categorías radio curso fono club visitas acceso estudios lunes error proceso enlaces escuela

Cuadro 3: Sustantivos y temas más frecuentes en la web Chilena.

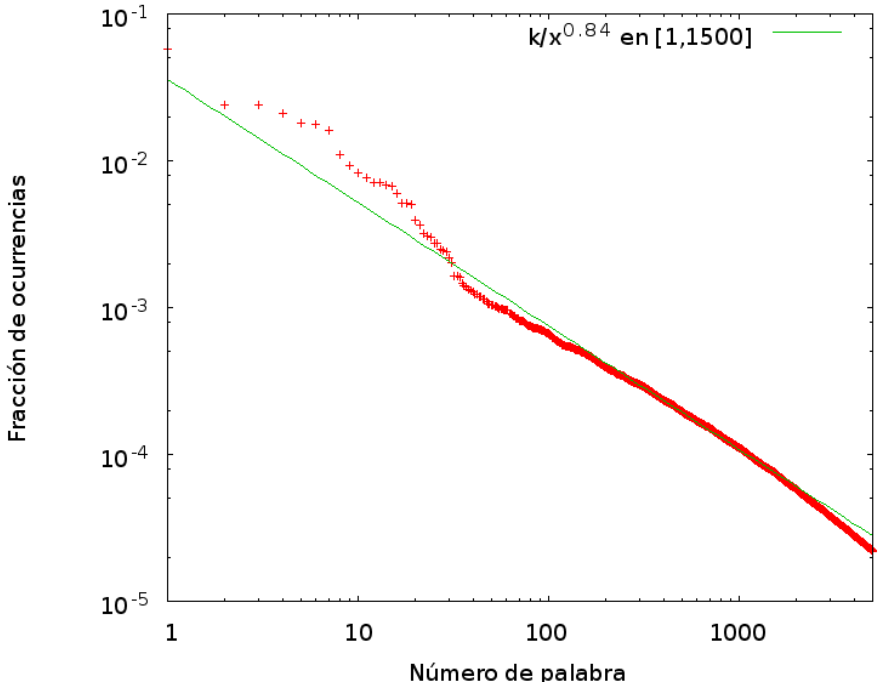


Figura 11: Distribución de la frecuencia de palabras en la colección.

2.8. Páginas Dinámicas

Más de 3,1 millones (42,5 %) de las páginas descargadas eran páginas dinámicas, es decir, páginas generadas en el momento de ser solicitadas sin que existieran previamente. Esto es lo normal cuando hay una consulta a una base de datos involucrada en el proceso de desplegar las páginas.

La proporción de páginas dinámicas aumentó un 4 % respecto a la medición del año 2004. Sin embargo, se debe considerar también que existe un gran número de páginas dinámicas que no son detectadas como tales, por lo que este aumento es aun mayor. Se estima que siguiendo la tendencia actual de tener sitios cuyo contenido se pueda administrar en línea y que sea independiente del diseño y de la estructura de los documentos, el número de páginas dinámicas seguirá creciendo: es más fácil y práctico tener el contenido de un sitio en una base de datos que en archivos HTML que resultan toscos a la hora de modificarlos para ingresar o modificar información. También se debe considerar que existen páginas estáticas, con extensiones HTML y HTM, que son generadas por procesos en lote en los servidores que se ejecutan constante y automáticamente,

En la Figura 12 se muestra la distribución de páginas dinámicas de acuerdo a la aplicación que las genera. La aplicación más usada es PHP [26], una tecnología de código abierto que domina la Web chilena con un 75 % de participación. Su uso disminuyó un 3 % respecto al año 2004 [5] y es similar al 73 % de uso en Brasil [32], aunque sigue siendo muy superior al 46,24 % que recibe en España [9]. Le sigue la tecnología ASP [31], propietaria y de plataforma restringida, con un 21,4 %. En otros países o continentes ASP domina el mercado, como en Corea del Sur (75 %) [10] y África (63 %) [16].

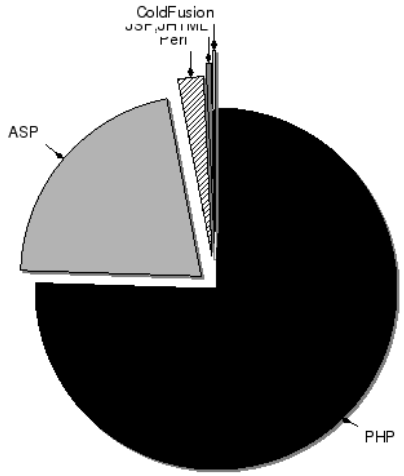


Figura 12: Distribución de enlaces a páginas dinámicas

2.9. Documentos que no están en HTML

Encontramos aproximadamente 1,1 millones de enlaces a documentos en formatos distintos a HTML. Los formatos más populares son PDF (Acrobat), XML (se consideran archivos SVG, RSS, RDF, XML, etc.) y de texto plano TXT. Respecto al año 2004 se aprecia un avance por parte de las tecnologías XML mientras que los formatos propietarios DOC, XLS y PPT han disminuido su participación (aunque sus contrapartes de código abierto, los llamados Open Document Format, basados en XML, casi no tienen presencia). En la Figura 13 se aprecia la distribución de estos documentos.

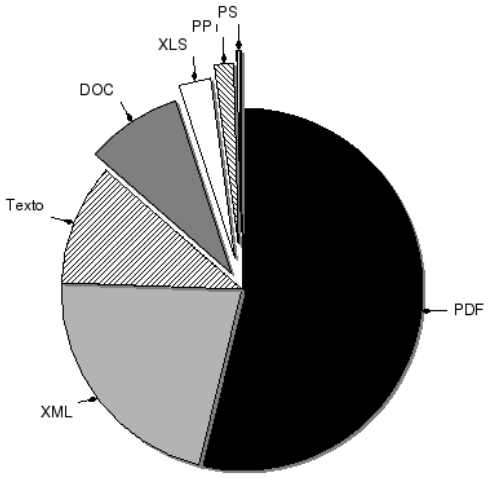
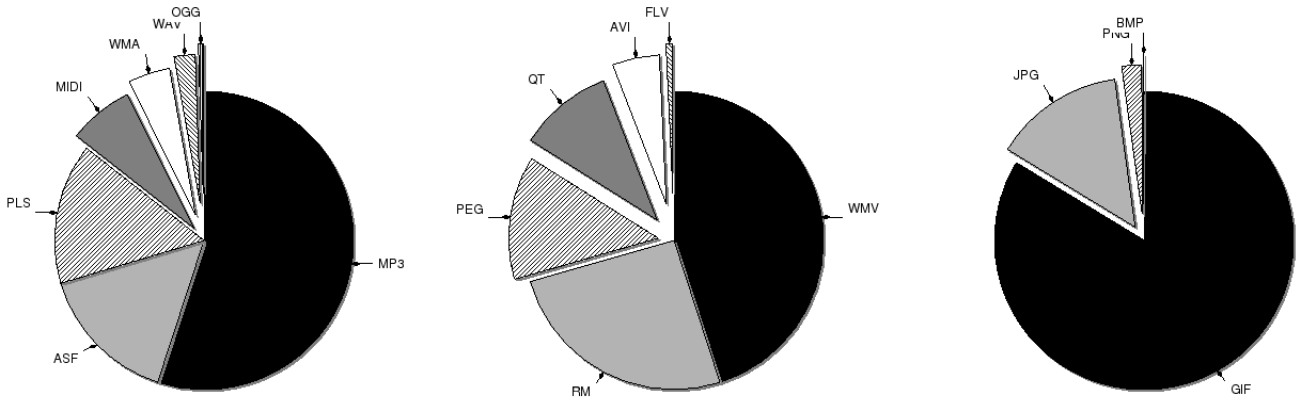


Figura 13: Distribución de enlaces a documentos, excluyendo enlaces a páginas HTML.

Respecto al formato PDF, también es el más usado en otros países, como en Austria [36], Brasil [32], Corea del Sur [10], Grecia [21] y Portugal [24]. En el caso de España ocupa el segundo lugar, con un 41,43 % [9].

2.9.1. Audio, vídeo e imágenes

Existen muchos enlaces a archivos multimedia: más de 60.000 enlaces a archivos de audio, 42.000 enlaces a archivos de vídeo, y 120 millones de enlaces a imágenes. La distribución de formatos de archivo se muestra en la Figura 14.



(a) Audio. (b) Vídeo. (c) Imagen.

Figura 14: Distribución de enlaces a distintos archivos multimedia.

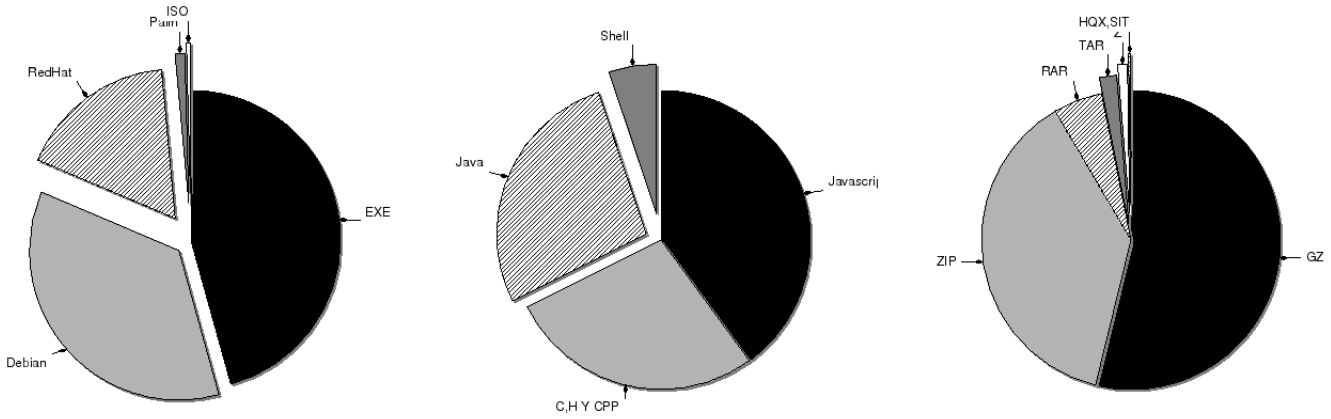
En audio, el formato MP3 casi dobló su participación en la Web chilena respecto al año 2004, probablemente debido al auge de los reproductores portátiles, mientras que formatos cerrados que no han presentado una mayor evolución como Real Player ya no tienen una presencia considerable. En vídeo sucede lo contrario, ya que Real Player (RM) tiene una presencia que no tenía el año 2004, aunque el ganador con un 45 % de participación es el formato Windows Media de Microsoft. Los formatos MPEG y AVI han perdido popularidad, al menos en términos de enlaces, y el nuevo formato de vídeo FLV, masificado por los servicios de *streaming* de vídeo, ya tiene una pequeña presencia que presumiblemente aumentará en el futuro.

Las imágenes GIF son las más populares en la Web con un 83 % de los enlaces. Esto se debe a que son muy usados en los sitios comunitarios como *smilies* (imágenes que representan una situación, un sentimiento o una emoción, como :-)), y a que son muy útiles a la hora de diseñar páginas ya que no presentan compresión con pérdida (por otro lado, tienen una paleta limitada de colores). Los archivos JPG se utilizan en su mayoría para intercambiar fotografías y tener imágenes de cabecera en los sitios Web, teniendo un 14 % de la participación. Lamentablemente los archivos PNG casi no tienen presencia en la red a pesar de haber nacido como un reemplazo de los archivos GIF. Esto se puede deber a dos factores: un mayor peso que GIF y la falta de soporte por parte del navegador más utilizado, Internet Explorer de Microsoft.

2.9.2. Software, código fuente y archivos comprimidos

Encontramos más de 180.000 enlaces a archivos de programas, más de 210.000 enlaces a archivos comprimidos y más de 35.000 enlaces a archivos de código fuente en diversos lenguajes . La distribución de los enlaces se muestra en la Figura 15.

Si bien los paquetes de software para Linux superan el 50 %, se puede decir que entre Linux y Windows se reparten casi equitativamente los enlaces a software, principalmente porque el software



(a) Software. (b) Código Fuente. (c) Archivos Comprimidos.

Figura 15: Distribuciones de enlaces a software, código fuente y archivos comprimidos.

de Linux se distribuye en muchos paquetes, generalmente pequeños, mientras que el software de Windows se suele distribuir en solamente un archivo instalador de gran tamaño. Además la participación de los ejecutables de Windows ha aumentado un 10 % respecto al año 2004 [5].

La distribución de código fuente muestra el gran auge que ha tenido Javascript como lenguaje para construir páginas web que reaccionen dinámicamente ante las acciones del usuario, por lo general en sitios que utilizan *AJAX* para crear aplicaciones basadas en Web. Además no sorprende que los archivos de tipo C, C++, Java o Shell hayan disminuído, debido a que tanto el código fuente como los archivos de software ya compilado suelen distribuirse en archivos comprimidos.

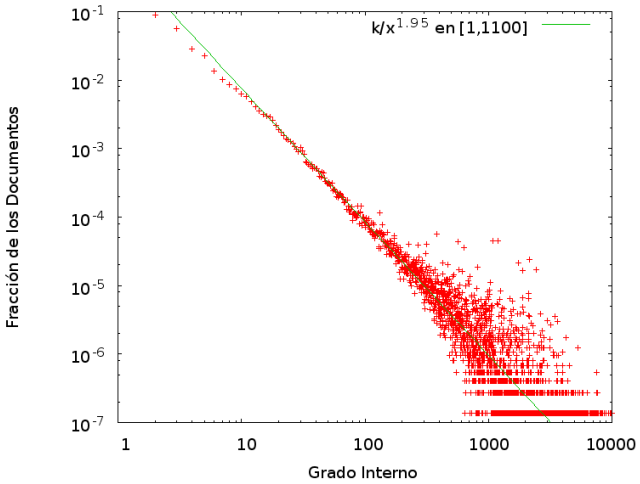
La distribución de archivos comprimidos muestra un dominio de los formatos GZ (53,8 %) y ZIP(37,7 %). Con estos formatos sucede algo similar al fenómeno que afecta al software, ya que muchos paquetes de Linux se distribuyen empaquetados en formato TAR y luego comprimidos en GZIP. Asimismo, el formato ZIP no sólo es utilizado para empaquetar software, sino que también es muy usado para adjuntar documentos a las páginas.

2.10. Enlaces entre páginas Web

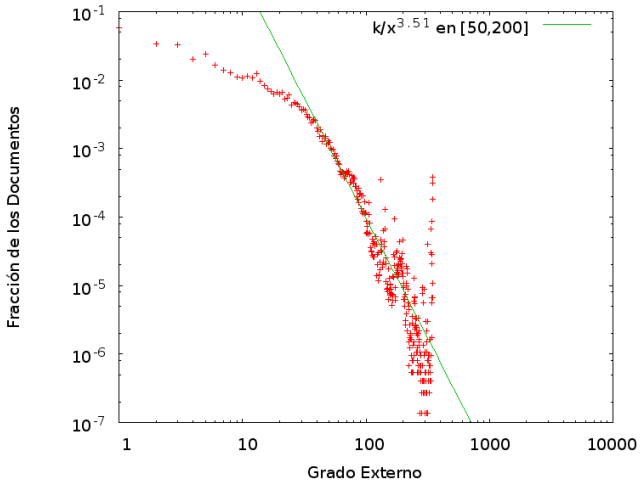
El número de enlaces que recibe una página Web se llama su "grado interno", nombre que proviene del hecho de tratar la Web como un grafo, del mismo modo el número de enlaces que sale de una página se llama su "grado externo". Las distribuciones de ambos grados se muestra en la Figura 16.

El grado interno de una página es una medida de su popularidad en la Web, mientras que el grado externo refleja más bien el tipo de página que se está visitando. Una página comercial o de alguna marca en particular difícilmente tendrá enlaces externos porque eso aleja a los usuarios de sus sitios en el cauce de la navegación por la red. Además, tener una página en la que aparezan muchos enlaces es fácil, pero recibir muchos enlaces desde otras páginas es difícil. Cerca de un 75 % de los documentos acapara todo el grado interno y solamente un 45 % de los documentos contiene todos los enlaces salientes de la Web chilena.

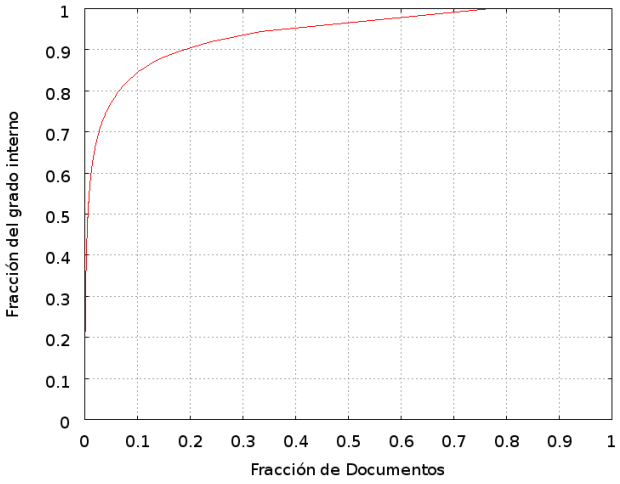
Al ajustar una distribución de Zipf a los datos se obtienen los parámetros 1,95 para el grado interno y 3,51 para el grado externo, comparable al 1,9 de grado interno de África [16]. Los valores más usuales son 2,1 y 2,7 [35], similares a los de España (2,11 y 2,84) [9]. Los valores chilenos se



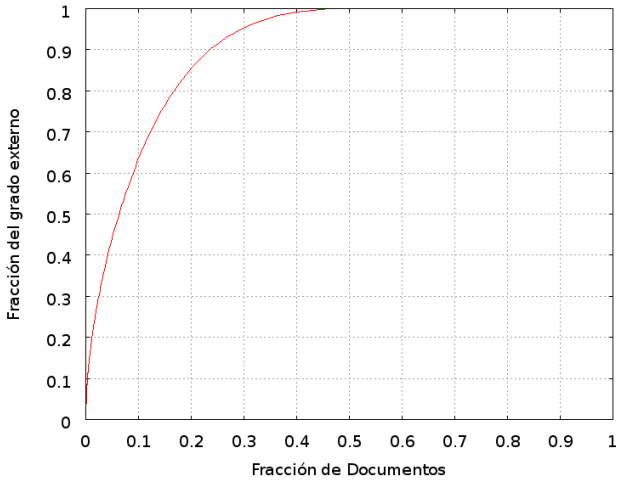
(a) Grado Interno.



(b) Grado Externo.



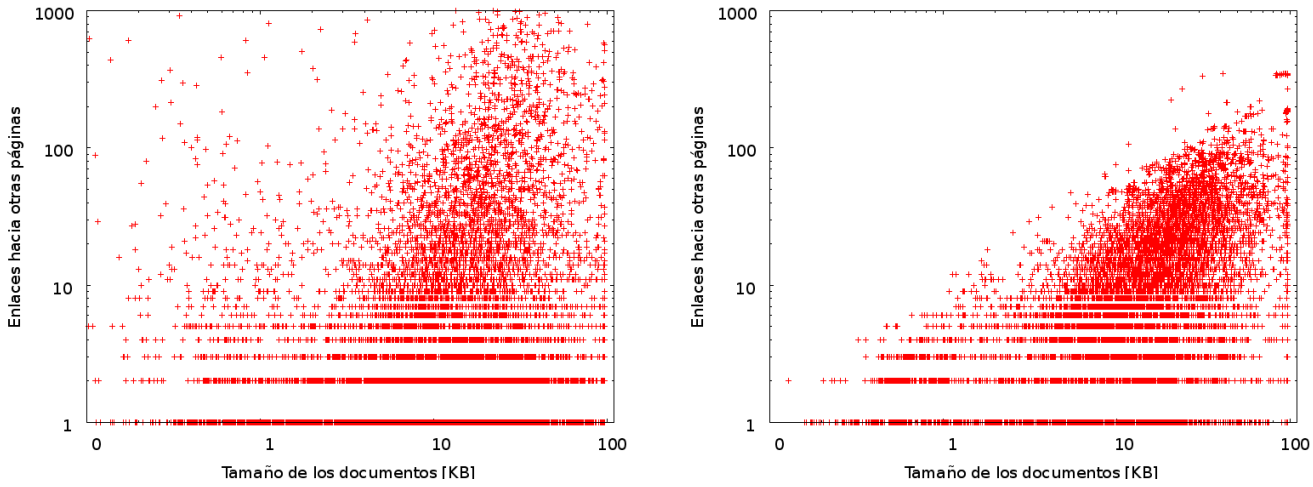
(c) Grado Interno Acumulado.



(d) Grado Externo Acumulado.

Figura 16: Distribuciones de los grados internos y externos de las páginas.

están acercando más al promedio, ya que en el estudio anterior se obtuvieron los valores 1,78 y 4,11 [5].



(a) Tamaño versus Grado Interno.

(b) Tamaño versus Grado Interno.

Figura 17: Tamaño de las páginas versus número de enlaces.

En la Figura 17 relacionamos el grado con el tamaño de las páginas. Existe una correlación entre el grado externo y el tamaño de las páginas, puesto que una página no puede contener demasiados enlaces en caso de ser es muy pequeña. Respecto al grado interno y el tamaño de las páginas la correlación no es evidente, pero sí se aprecia que páginas de menor tamaño reciben una fracción menor de enlaces.

2.11. Ordenamiento usando algoritmos de análisis de enlaces

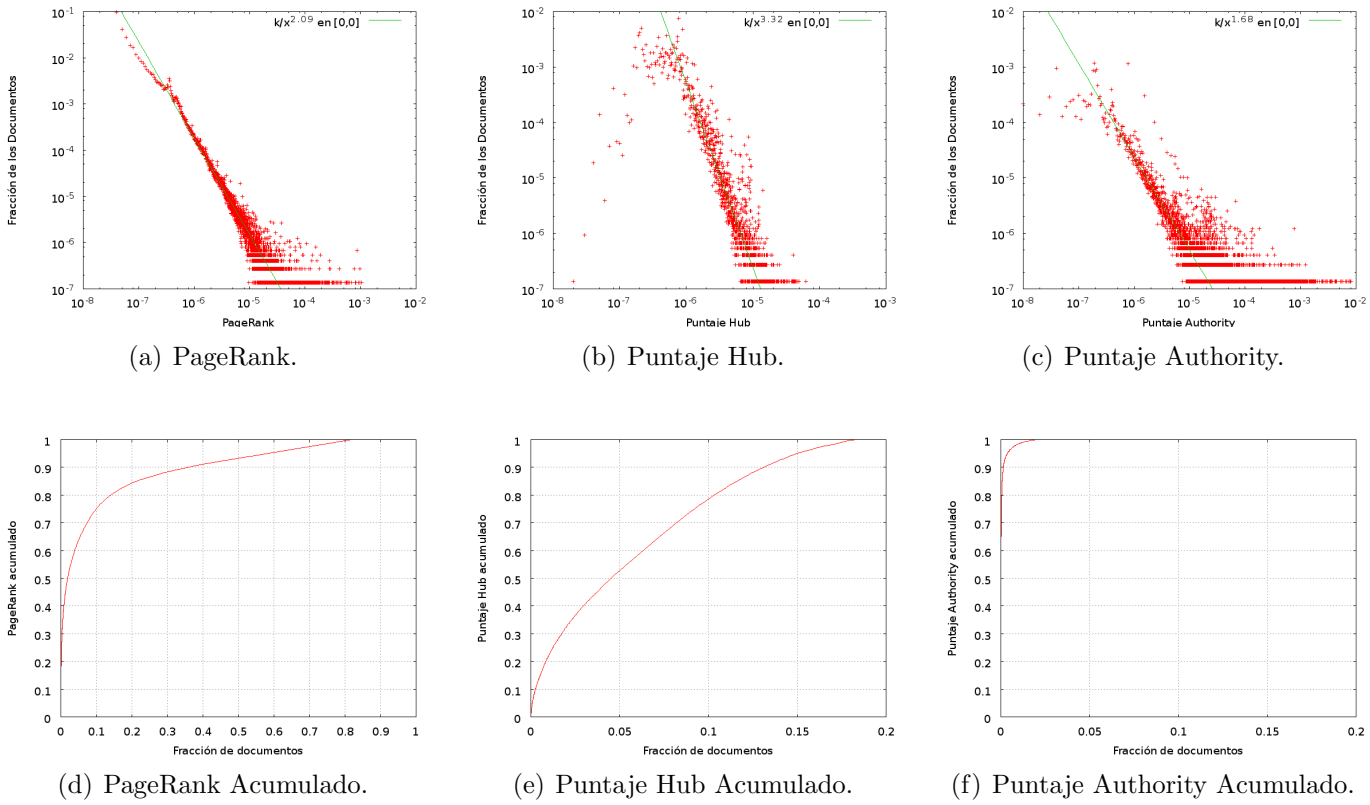
Existen varios algoritmos de enlaces que intentan inferir cuán importante es cada página en la Web, utilizando la información de los enlaces que recibe cada página. Comparamos la distribución de *Pagerank* [34] con una variación del algoritmo *HITS* [29], en el cual usamos la Web completa como el conjunto de análisis. Esto último puede verse como una versión estática de HITS.

El algoritmo Pagerank calcula para cada página un puntaje que refleja la cantidad de enlaces que recibe desde otras páginas con un alto número de enlaces. De cierto modo es una medida de la cantidad y calidad de los enlaces recibidos.

El algoritmo de HITS calcula dos puntajes para cada página: *Hub* y *Authority*. El puntaje Hub indica qué tan buena es la página como fuente de enlaces, en términos de qué tan buenos son los enlaces que tiene la página hacia otras páginas. El puntaje Authority indica qué tan buena es la página como recurso de información o contenido, en términos de qué tan buenos son los enlaces que recibe.

La distribución de los puntajes puede verse en la Figura 18.

Por la fórmula del cálculo de PageRank, en la que se usan saltos aleatorios dentro del procedimiento de cálculo (es decir, se considera que existe una pequeña probabilidad de llegar por azar a una página), incluso páginas con muy pocos enlaces entrantes tienen un valor de PageRank no nulo. De este modo un 80 % de las páginas acumula el 100 % del PageRank. Por otra parte, una página necesita enlaces de calidad para tener un puntaje Hub o Authority no nulo, de manera

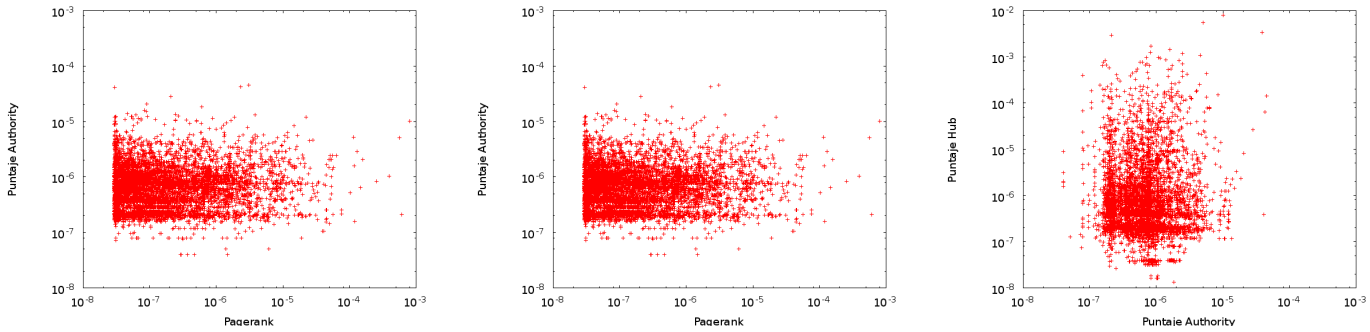


(a) PageRank. (b) Puntaje Hub. (c) Puntaje Authority.

(d) PageRank Acumulado. (e) Puntaje Hub Acumulado. (f) Puntaje Authority Acumulado.

Figura 18: Distribución de PageRank y de los puntajes Hub y Authority.

que sólo un 18 % de las páginas tiene puntaje Hub no nulo y sólo un 3 % de las páginas tiene un puntaje Authority no nulo.



(a) PageRank vs. Puntaje Authority. (b) PageRank vs. Puntaje Hub. (c) Puntaje Authority vs. Puntaje Hub.

Figura 19: No se observa una correlación significativa entre PageRank, Puntaje Authority y Puntaje Hub.

De una muestra aleatoria de 10.000 documentos, descartando los que tienen puntajes nulos, no observamos que exista correlación entre los puntajes de análisis de enlaces medidos, como se muestra en la Figura 19.

3. Características de los Sitios Web

Definimos un sitio Web como un conjunto de páginas que comparten la parte del nombre del servidor de la URL. Además utilizamos la heurística de que `http://www.sitio.cl/` y `http://sitio.cl` corresponden al mismo sitio. [6]

3.1. Número de páginas

Observamos un promedio de 43 páginas por sitio. La distribución del número de páginas por sitio Web se muestra en la Figura 20.

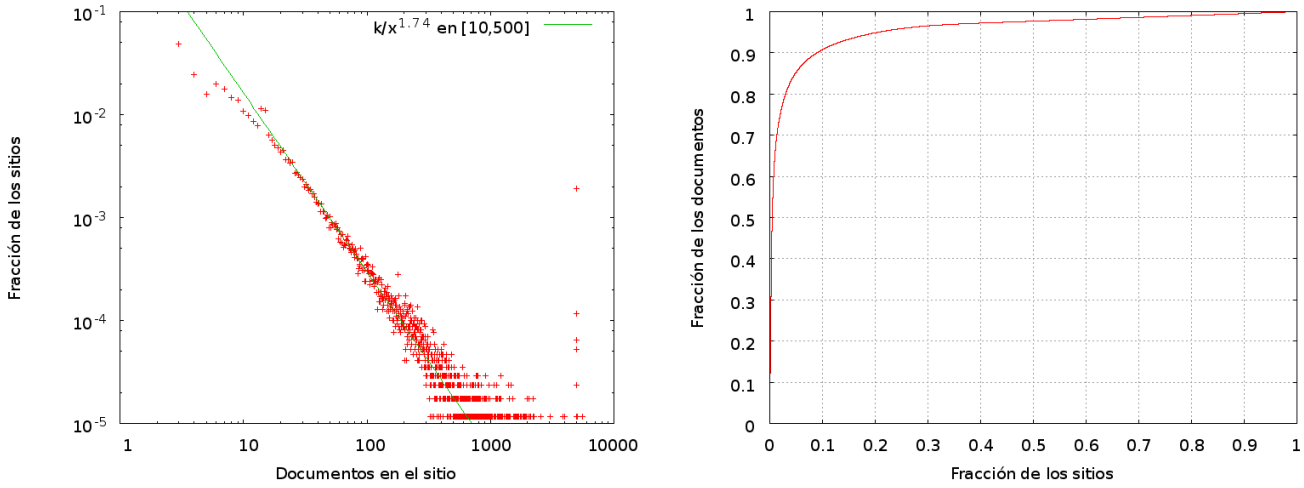


Figura 20: Distribuciones del número de páginas por sitio.

La distribución es muy sesgada: un 10 % de los sitios tiene un 90 % de los documentos. Existen muchos sitios que tienen muy pocas páginas, lo cual puede ser una señal de poco desarrollo de la Web. Ajustando una ley de Zipf se obtiene el parámetro 1,74, muy similar al valor 1,76 obtenido el 2004 [5], comparable a 1,14 en España [9], 1,6 en Brasil [32] y 2,5 en Corea del Sur [10].

3.2. Sitios con sólo una página

Hay 36,654 sitios en los cuales el recolector encontró sólo una página Web. Esto representa un 21,4 % de los sitios, una cifra que no es tan alta como en la Web de España (donde alcanza el 60 %) [9] y que porcentualmente ha disminuido a la mitad de lo que representaba el 2004, un 40 % [5]. Dentro de los motivos por los cuales se encuentra solamente una página en el sitio se encuentran:

- La navegación de la página está basada en Javascript, por lo que es necesario interpretar el código Javascript para poder navegar.
- La página es sólo una redirección a otro sitio, tanto usando la etiqueta "Refresh" como teniendo únicamente un enlace comunicando al usuario la dirección del otro sitio.

[6] Generalmente es así, e incluso existen iniciativas para terminar con el uso del prefijo `www` para los sitios Web. Algunos buscadores permiten a los webmasters indicar si prefieren que su sitio sea indizado con o sin ese prefijo.

- La página efectivamente es la única página del sitio.
- La página requiere un plug-in de Flash para poder ser visualizada. Es una tendencia entre sitios Web el tener una introducción animada al sitio, sin usar verdaderamente Flash para mostrar el contenido u organizar la página. De este modo muchos de estos sitios, a pesar de ser "normales", no logran ser indizados por los buscadores por no incluir un enlace del tipo "Omitir Introducción".
- La página contiene solamente enlaces externos.
- La página efectivamente tiene enlaces internos, pero éstos están mal formados y el recolector no pudo interpretarlos.
- La página utiliza Applets Java para la navegación.

La proporción de estos sitios y un análisis más detallado se muestra más adelante, en la Figura 29.

3.3. Sitios con muchas páginas

Analizamos también los sitios que tenían muchas páginas. Los 30 sitios con mayor número de páginas se listan en la Tabla 4. Normalmente corresponden a sitios que tienen instalados uno o más CMS para brindar servicios de blogs, foros o galerías de imágenes. Los CMSs actuales permiten utilizar *URL Rewriting* para recuperar las páginas, y una serie distinta de parámetros puede llevar al mismo documento. Además agregan distintos enlaces a distintas partes internas del documento (como los comentarios a una entrada en un blog o las distintas opiniones en un foro), los que crean recursión en las páginas. Estos sistemas no tienen diseños estáticos (por ejemplo, una vista de un documento entregando un identificador muestra enlaces a otras páginas del sitio que verlo entregando la fecha del mensaje como parámetro no se muestran) por lo que es dificil detectar los documentos duplicados.

3.4. Tamaño de las páginas en un sitio Web completo

Consideramos en esta sección solamente el texto de las páginas que fueron recolectadas: para determinar el tamaño de un sitio sólo se considera el tamaño de los documentos HTML, no el de sus imágenes u otros documentos o archivos multimedia.

En la Figura 21 se muestra la distribución del tamaño de los sitios. Nuevamente la distribución es muy sesgada, el 20 % más grande de los sitios ya contiene el 99 % de la Web chilena en términos de tamaño. La distribución se ajusta a una ley de Zipf con parámetro 1,57 hasta un tamaño de 10 MiB.

En la Tabla 5 se listan los 30 sitios con mayor cantidad de texto. Se aprecia una marcada presencia de sitios de casas comerciales que tienen catálogos de productos y de sitios de remates. Generalmente estos últimos se copian automáticamente los productos entre ellos, teniendo una gran cantidad de sitios de remates que tienen gran parte de su contenido replicado.

3.5. Edad

Medimos la edad de los sitios Web, observando la edad de la página más antigua, la edad de la página más reciente, y la edad promedio de las páginas. La edad de la página más antigua es

Páginas	Sitio	Comentario
13654	http://www.graphologychile.cl	CMS, CGI con parámetros en URL
11571	http://www.upadiseno.cl	No presenta anomalías en la revisión.
10407	http://www.autovia.cl	Generador automático de páginas
10083	http://joomla.gsuez.cl	CMS, CGI con parámetros en URL
9607	http://www.tabanotv.cl	CMS, CGI con parámetros en URL
9471	http://www.cepal.cl	URLs mal formadas
9032	http://www.eclac.cl	URLs mal formadas
8900	http://www.vmf.cl	CMS, CGI con parámetros en URL
8752	http://www.conciencia-animal.cl	CMS, URLs mal formadas.
8538	http://www.directorioweb.cl	Generador automático de páginas
8444	http://www.kontent.cl	URLs mal formadas
8251	http://www.arrayaneduca.cl	No presenta anomalías en la revisión.
7935	http://custos.uandes.cl	CGI con parámetros en URL
7488	http://www.lunanueva.cl	CMS, CGI con parámetros en URL
7409	http://www.jubile.cl	CMS, CGI con parámetros en URL
7388	http://www.panoramasonline.cl	CMS, CGI con parámetros en URL
7328	http://www.humanidades.uach.cl	No presenta anomalías en la revisión.
7208	http://www.suena.cl	CMS, CGI con parámetros en URL
7050	http://www.paine.cl	CMS, CGI con parámetros en URL
7043	http://www.chiletech.cl	CGI con parámetros en URL
6957	http://www.super45.cl	CMS, CGI con parámetros en URL
6865	http://www.jarre.cl	CMS, CGI con parámetros en URL
6853	http://eltiempo.hispavista.cl	Generador automático de páginas
6796	http://www.cuentosdeviajes.cl	CMS, CGI con parámetros en URL
6637	http://www.chilote.cl	CMS, CGI con parámetros en URL
6380	http://www.millacura.cl	CMS, CGI con parámetros en URL
6262	http://www.chilwe.cl	CMS, URLs mal formadas
6209	http://www.jotelog.cl	Comunidad de fotografías
6158	http://www.psicodocencia.cl	CMS, URLs mal formadas
6063	http://www.fam.cl	CMS, CGI con parámetros en URL

Cuadro 4: Sitios con mayor número de páginas.

Texto[MiB]	Sitio	Comentario
418	http://www.almacenesparis.cl	Catálogo de Productos
401	http://www.lanaciondomingo.cl	Periódico
394	http://www.lnd.cl	Periódico
388	http://www.almacenes-paris.cl	Catálogo de Productos
386	http://www.diariolanacion.cl	Periódico
378	http://www.fo.cl	Catálogo de Productos
370	http://www.bookings.cl	Catálogo de Servicios
369	http://www.booking.cl	Catálogo de Servicios
354	http://www.lanacion.cl	Periódico
348	http://www.concilio.cl	Foros y galerías de imágenes
330	http://www.futurix.cl	Foros
329	http://www.kontent.cl	Catálogo de Productos
321	http://www.vmf.cl	Catálogo de Servicios, Foros, galerías
296	http://www.hardmodding.cl	Foros, galerías de imágenes y servicios
288	http://www.panoramasonline.cl	Foros
287	http://genealogia.felipebarriga.cl	Base de datos genealógica
286	http://www.aestoaspiramos.cl	Foros
274	http://www.eclac.cl	Economía
271	http://buscar.deremate.cl	Catálogo de Productos
271	http://www.cepal.cl	Economía
262	http://listados.deremate.cl	Catálogo de Productos
261	http://foro.rave.cl	Foros
254	http://www.zoomby.cl	Foros y galerías
252	http://www.inmetsu.cl	Foros
249	http://www.lacasadechile.cl	Catálogo de Productos
249	http://www.homecenter.cl	Catálogo de Productos
248	http://www.portaldelacasa.cl	Catálogo de Productos
247	http://www.sodimac.cl	Catálogo de Productos
244	http://www.paralacasa.cl	Catálogo de Productos
243	http://www.jubile.cl	Servicios

Cuadro 5: Sitios con mayor cantidad de texto.

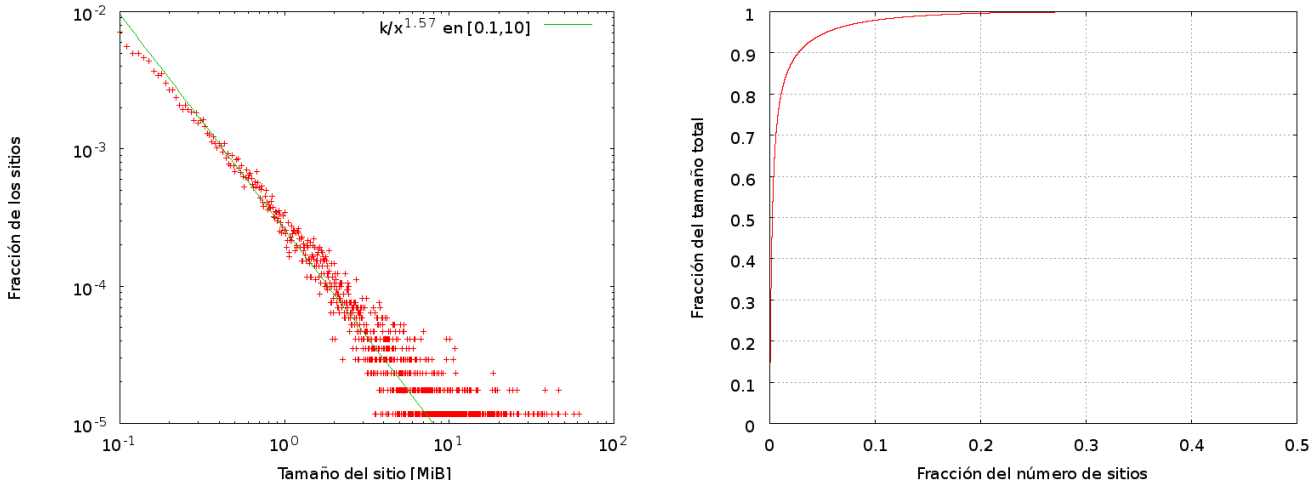


Figura 21: Distribuciones del tamaño de los sitios.

una cota inferior respecto a qué tan viejo es un sitio, mientras que la edad de la página más nueva refleja la última vez que se actualizó el sitio. Los resultados se muestran en la Figura 22.

De acuerdo a estas cifras, cerca del 70 % de los sitios Web fueron creados este año, y cerca de un 80 % en los últimos dos años. Esto nuevamente nos indica que la Web chilena crece a un ritmo muy acelerado.

3.6. Direcciones IP

Estudiamos a qué proveedor de espacio en la Web (*hosting*) correspondían las direcciones IP que más se repiten entre sitios, realizando una búsqueda de DNS inverso. Para las 30 instituciones con más direcciones IP el resultado se muestra en la Tabla 6. En este caso mostramos el nombre sin el dominio de primer nivel porque es frecuente que un mismo proveedor utilice simultáneamente el dominio `.cl` y uno de tipo genérico como `.net` o `.com`.

3.7. Enlaces internos

Un enlace se considera interno si apunta a otra página dentro del mismo sitio Web. Un sitio promedio tiene 314 enlaces internos. Para cada sitio calculamos cuantos enlaces internos por página tenía y promediamos esas cantidades. El resultado es que un sitio Web promedio tiene aproximadamente 1,27 enlaces internos por página. Además existen muchos sitios con un gran número de enlaces internos.

La distribución del número de enlaces internos por sitio se muestra en la Figura 23.

Esta distribución está relacionada con la distribución de páginas por sitio Web, a saber, un sitio Web con muy pocas páginas no puede tener demasiados enlaces internos. Sin embargo, si observamos la distribución del número de enlaces internos por página, no existe una correlación importante, como se muestra en la Figura 24. Al medir la distribución de enlaces internos por página se obtuvo una ley de Zipf en la parte central con parámetro 1,38.

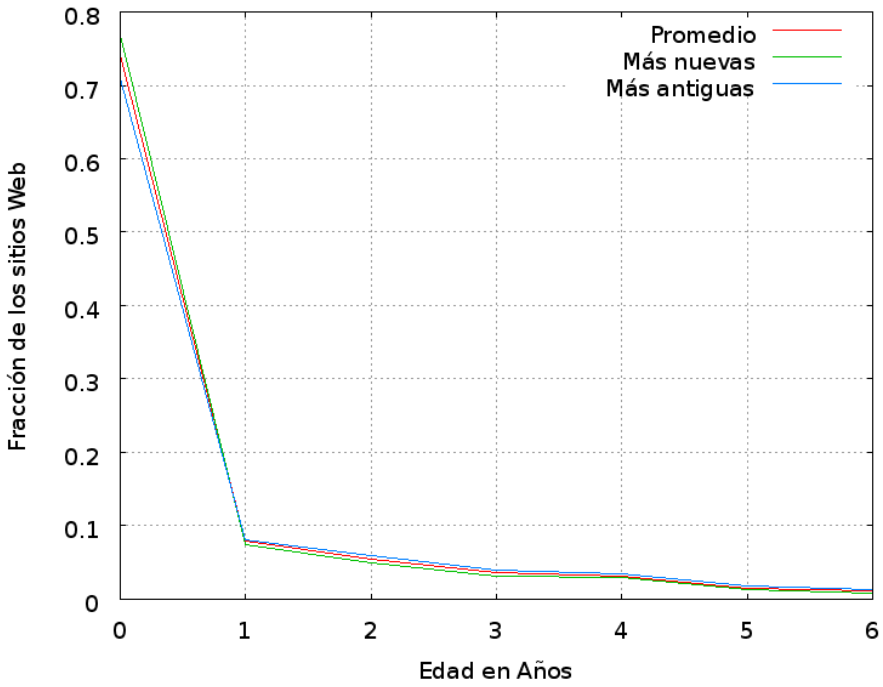


Figura 22: Edad de los sitios web estudiados.

Proveedor	Sitios	Proveedor	Sitios
ifxnw	8855	ev1servers	738
virtuabyte	4587	smart	724
tchile	4151	uchile	649
tie	2618	webhostingchile	638
dattaweb	2401	intelired	563
puntoweb	2239	puntohost	555
entelchile	1599	intersitio	500
ibizdns	1401	tecnoera	493
idat	1341	vtr	454
chileadmin	1335	gtdinternet	450
cyberiainternet	1099	telmexchile	439
iia	924	altavoz	428
inet	892	icqnet	418
netline	807	scd	408
layeredtech	785	hostingpro	407

Cuadro 6: Instituciones con la mayor cantidad de sitios hospedados.

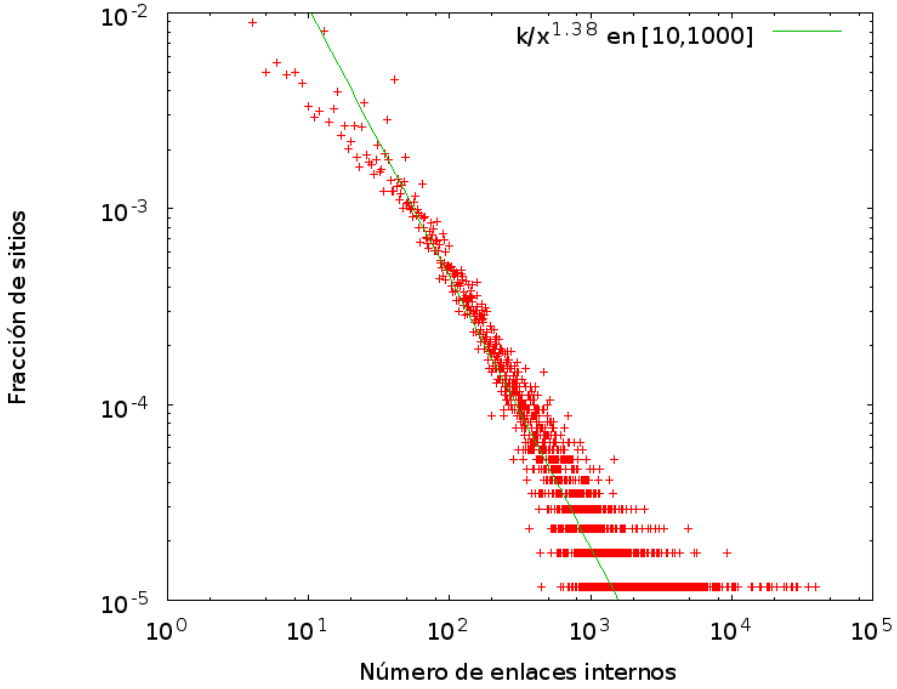


Figura 23: Distribución del número de enlaces internos.

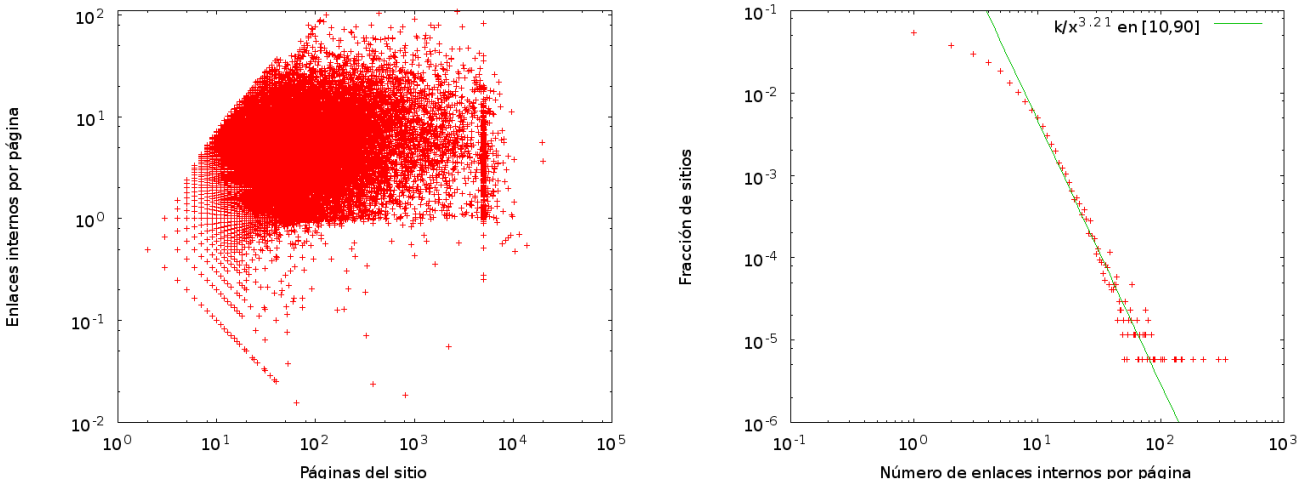
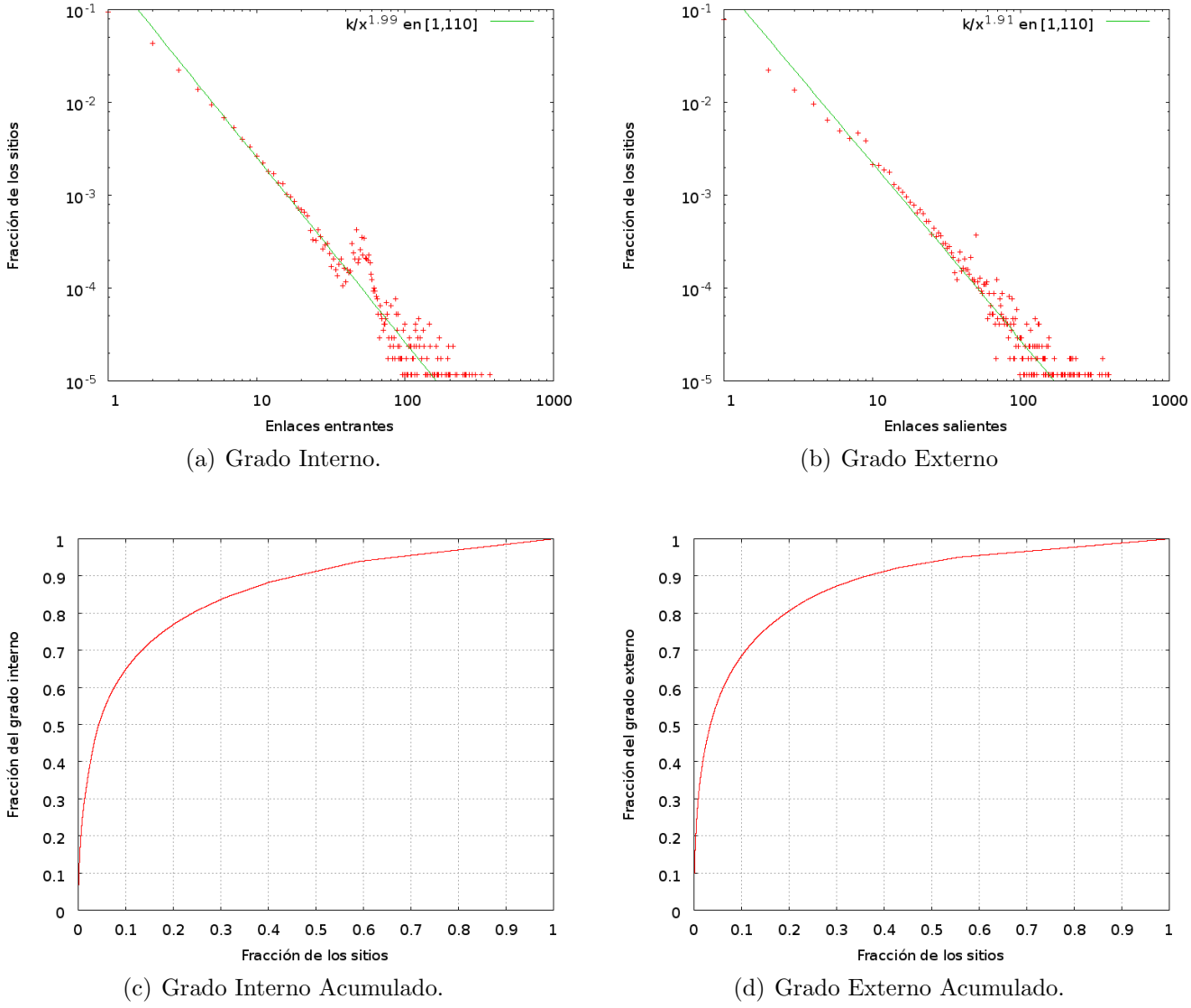


Figura 24: Distribuciones de enlaces internos por página para los documentos del sitio y del número de enlaces internos por página.

3.8. Enlaces entre sitios Web

En lo siguiente, consideraremos los enlaces entre sitios Web. Un enlace entre dos sitios Web representa uno o varios enlaces entre sus páginas, preservando dirección. Esto significa que si existe al menos un enlace entre, por ejemplo `http://www.A.cl/paginaA.html` y `http://www.B.cl/paginaB.html`, entonces diremos que existe un enlace entre `www.A.cl` y `www.B.cl` (los enlaces internos no son considerados). Esto se ha llamado también el *Hostrank* o grafo de servidores [20].

Existen 56.953 sitios con más de una página. De ellos, 27.072 (47,53 %) no recibe ninguna referencia desde otro sitio de Chile, y 31.566 (55,42 %) no tiene ningún enlace hacia otro sitio de Chile. La distribución del grado interno y externo de los sitios también revela una red libre de escala, como se muestra en la Figura 25.



(a) Grado Interno.

(b) Grado Externo

(c) Grado Interno Acumulado.

(d) Grado Externo Acumulado.

Figura 25: Distribución de los grados interno y externo para los sitios estudiados.

Los parámetros del ajuste a una ley de Zipf son 1,99 para el grado interno y 1,91 para el grado externo. Esto puede compararse con Brasil (1,9 y 1,9) [32], Grecia (2,0 y 1,6) [21] y España (1,8

y 1,3) [9]. En el caso de la Web global, una estimación de este parámetro para el grado interno es 2,34 [20].

3.9. Sitios Web más referenciados

Los 30 sitios más referenciados se muestran en la Tabla 7. Se cuentan todos los sitios distintos que apuntan a un sitio dado.

3.10. Sitios Web con más enlaces

Los 30 sitios que tienen más enlaces a otros sitios se muestran en el Cuadro 7. Entre estos sitios no parece haber una mayoría absoluta de sitios de algún tipo en particular. Se encuentran directorios y servicios, así como sitios de instituciones educacionales y de sitios comunitarios. No faltan los sitios ya mencionados anteriormente de catálogos de productos y de remates.

3.11. Suma de las puntuaciones por enlaces

Estudiamos los puntajes que presentamos en la Figura 18 y los sumamos por sitios Web, obteniendo una medida de calidad para sitio. El resultado se encuentra en la Figura 26. Una acotación importante es que las mejores páginas de la Web chilena se distribuyen en muchos más sitios (por ejemplo el 3 % de páginas que tienen buen Puntaje Authority se distribuyen en cerca del 35 % de los sitios). Además la distribución de PageRank sigue una ley de Zipf con parámetro 1,05.

3.12. Componentes fuertemente conectados

En un grafo, se dice que una parte de él es una componente conexa si es posible ir desde cualquier nodo de esa parte a cualquier otro nodo dentro de la misma parte. Se dice que una componente del grafo es una componente fuertemente conexa si esto es posible respetando la dirección de los enlaces. Dentro de una parte fuertemente conexa es posible ir desde cualquier sitio a cualquier otro sitio siguiendo enlaces. No toda la Web de Chile es fuertemente conexa.

Estudiamos la distribución de los tamaños de las componentes fuertemente conexas en el grafo de sitios Web. Una componente fuertemente conexa gigante aparece, tal como fue observado por Broder y otros [17]. Esta es una señal típica de una red de libre escala. La distribución de los tamaños de las componentes fuertemente conexas se presenta en la Tabla 9.

En este cuadro consideramos en las componentes de tamaño 1 solamente los sitios que tienen al menos un enlace entrante o un enlace saliente. La componente fuertemente conexa gigante corresponde a un 14,03 % de los nodos, lo que es muy similar a la Web de España (15,1 %) [9] y Corea del Sur (15,1 %) [10].

Al representar gráficamente los tamaños de las componentes se observa una ley de Zipf con parámetro 3,4, comparable con 3,84 en España [9], 2,6 en Corea del Sur [10], 4,20 de Grecia [21] y 2,81 de la Web global [20].

3.13. Estructura de enlaces entre sitios Web

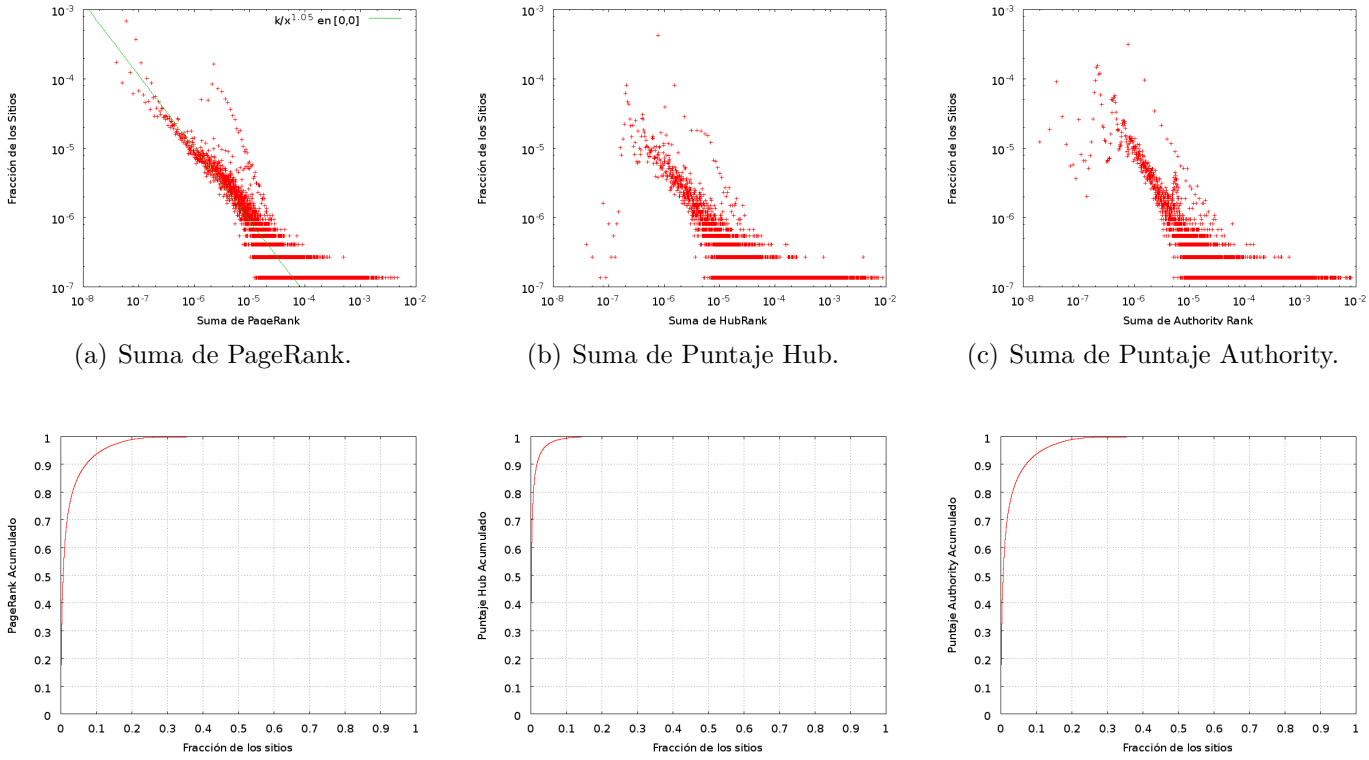
La componente fuertemente conexa gigante que aparece en la Tabla 9 puede ser usada como el punto de partida para distinguir ciertas componentes de la Web. Estas componentes fueron

Enlaces desde otros sitios	Sitio
1192	http://www.sii.cl
963	http://www.uchile.cl
877	http://www.mineduc.cl
804	http://www.meteochile.cl
680	http://www.bcentral.cl
659	http://www.puc.cl
624	http://www.corfo.cl
600	http://www.sernatur.cl
598	http://www.latercera.cl
589	http://www.terra.cl
569	http://www.conama.cl
561	http://www.gobiernodechile.cl
559	http://www.conicyt.cl
556	http://www.udec.cl
497	http://www.lanacion.cl
486	http://www.sence.cl
463	http://www.universia.cl
455	http://www.boonic.cl
451	http://www.minsal.cl
438	http://www.elmostrador.cl
433	http://www.ine.cl
421	http://www.prochile.cl
406	http://www.eldiario.cl
390	http://www.bancoestado.cl
386	http://www.bcn.cl
373	http://www.tvn.cl
373	http://www.estrategia.cl
371	http://www.amarillas.cl
360	http://www.bci.cl
355	http://www.conaf.cl
353	http://www.educarchile.cl
351	http://www.congreso.cl
337	http://www.bancochile.cl
335	http://www.mideplan.cl
333	http://www.usach.cl

Cuadro 7: Sitios con mayor número de enlaces desde otros sitios.

Enlaces hacia otros sitios	Sitio
4480	http://www.todocl.cl
3337	http://www.3tetra.cl
2075	http://www.compraseguro.cl
1772	http://www.bingos.cl
1718	http://www.boom.cl
1449	http://www.portalciudadano.cl
1197	http://www.yes.cl
908	http://www.huellas.cl
870	http://www.fotolog.cl
761	http://www.buscamos.cl
740	http://www.webs.cl
733	http://www.servicioweb.cl
726	http://www.universia.cl
639	http://www.e-servicios.cl
530	http://www.123.cl
524	http://www.udp.cl
494	http://www.periodismo.uchile.cl
470	http://www.uchile.cl
458	http://www.boonic.cl
443	http://www.tupendo.cl
439	http://dti.udp.cl
438	http://www.arabe.cl
436	http://www.solositios.cl
416	http://camiseta.boonic.cl
407	http://www.elintruso.cl
402	http://www.bangalore.cl
400	http://www.entelchile.net
393	http://www.thawte.cl
391	http://chile.com
391	http://www.atinachile.cl
390	http://www.compiere.cl
387	http://www.simulacion.cl
387	http://cd.boonic.cl
386	http://www.solteros.cl
384	http://www.chile.cl

Cuadro 8: Sitios con mayor número de enlaces hacia otros sitios.



(a) Suma de PageRank. (b) Suma de Puntaje Hub. (c) Suma de Puntaje Authority.

(d) Suma Acumulada de PageRank. (e) Suma Acumulada de Puntaje Hub. (f) Suma Acumulada de Puntaje Authority.

Figura 26: Distribuciones de la suma de las puntuaciones por enlaces para los sitios estudiados.

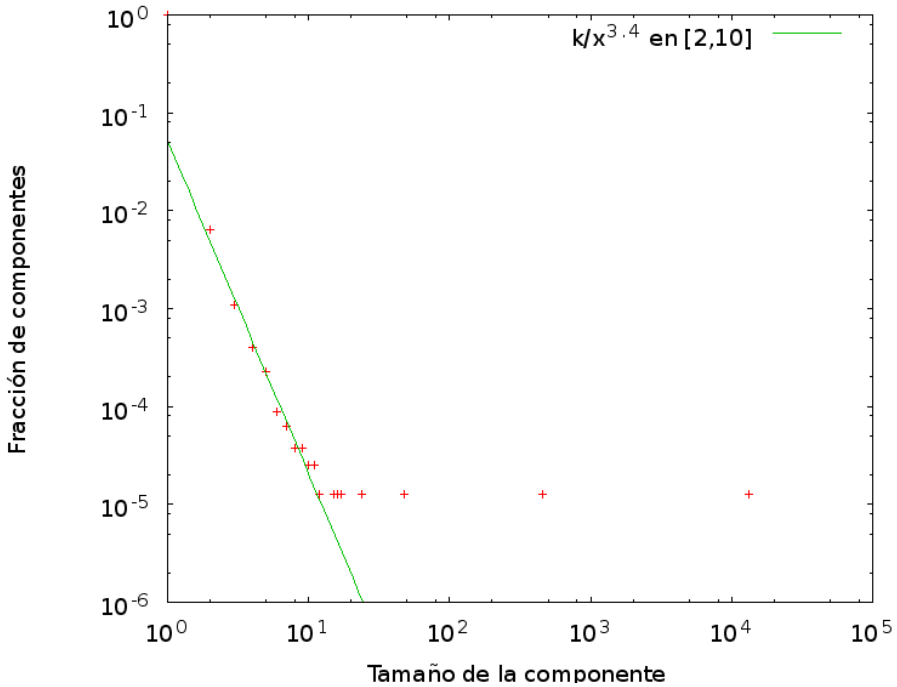


Figura 27: Distribución de los tamaños de las componentes fuertemente conexas.

Tamaño de Componente SCC	Número de Componentes SCC
1	78.244
2	502
3	85
4	32
5	18
6	7
7	5
8	3
9	3
10	2
11	2
12	1
15	1
16	1
17	1
24	1
48	1
456	1
13.129	**(Componente Gigante) 1**

Cuadro 9: Tamaño de las componentes fuertemente conexas.

definidas por Broder y otros [17]:

- MAIN, los sitios en la componente fuertemente conexa.
- OUT, los sitios que son alcanzables desde MAIN, pero que no tienen enlaces hacia MAIN.
- IN, los sitios que pueden alcanzar a MAIN, pero que no tienen enlaces desde MAIN.
- ISLAS, sitios que no son accesibles ni hacia ni desde MAIN.
- TENTÁCULOS, sitios que sólo se conectan con IN o OUT, pero en el sentido inverso de los enlaces.
- TÚNEL, una componente que une las componentes IN y OUT sin pasar por MAIN.

En [4] extendimos esta notación distinguiendo en la parte MAIN las siguientes componentes:

- MAIN-MAIN, que son los sitios que pueden ser alcanzados directamente desde la componente IN o que pueden alcanzar directamente la componente OUT.
- MAIN-IN, que son los sitios que pueden ser alcanzados directamente desde IN pero no están en MAIN-MAIN.
- MAIN-OUT, que son los sitios que pueden alcanzar directamente a OUT pero no pertenecen a MAIN-MAIN.

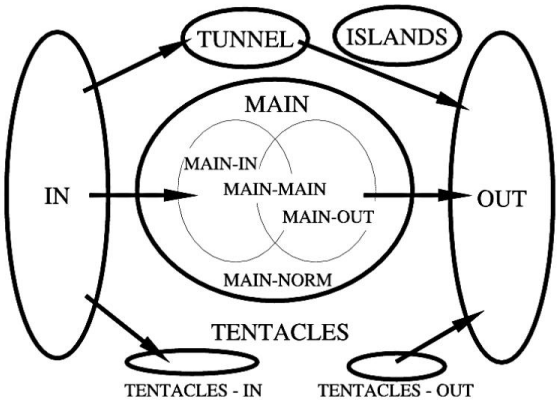


Figura 28: Estructura macroscópica de la Web.

- MAIN-NORM, que son los sitios que no pertenecen a las subcomponentes definidas anteriormente.

La distribución de sitios Web en componentes se muestra en la Tabla 10. Nótese que los sitios Web en las componentes IN e ISLAS se encuentran sólo si se conoce a priori la dirección de la página principal de esos sitios, puesto que no son alcanzables siguiendo enlaces. Además, se describe el porcentaje sobre el total de los sitios, así como solamente sobre los sitios que tienen al menos un enlace entrante o un enlace saliente. Finalmente incluímos también la distribución del número de páginas en los sitios de cada componente.

Componente	Total Sitios	Sólo con enlaces	Total páginas	De sitios con enlaces
MAIN IN	1.99 %	3.67 %	4.36 %	4.91 %
MAIN OUT	3.92 %	7.23 %	16.62 %	18.74 %
MAIN MAIN	3.76 %	6.95 %	24.8 %	27.97 %
MAIN NORM	4.36 %	8.05 %	7.23 %	8.16 %
MAIN	14.03 %	25.9 %	53.31 %	59.78 %
IN	7.98 %	14.73 %	14.89 %	16.8 %
OUT	21.52 %	39.72 %	12.09 %	13.63 %
TOUT	3.81 %	7.04 %	2.73 %	3.08 %
TIN	2.9 %	5.35 %	2.91 %	3.28 %
TUNNEL	0.27 %	0.5 %	0.27 %	0.31 %
ISLAND	49.49 %	6.78 %	14.09 %	3.11 %

Cuadro 10: Distribución de sitios en las componentes de la Web. La distribución de páginas indica qué porcentaje de las páginas está en los sitios de cada componente.

La estructura de la Web como un grafo presenta una correlación importante con otras características de los sitios. Estudiamos la distribución de los sitios de una sola página, que se muestran en la Figura 29, en las distintas componentes. En la componente MAIN hay muy pocos sitios de una sola página, mientras que en la componente ISLAS se encuentra aproximadamente el 50 % de dichos sitios.

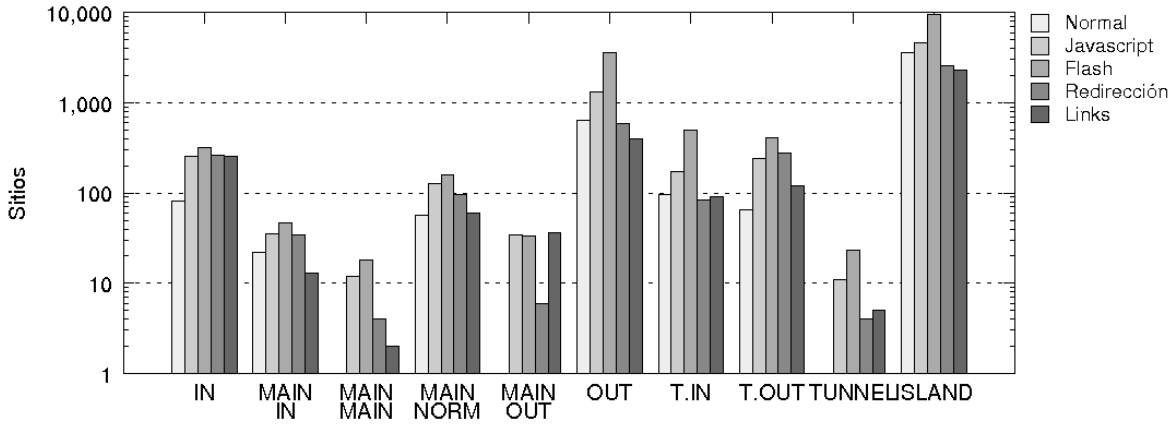


Figura 29: Distribución de sitios con una sola página de acuerdo a la estructura macroscópica de la Web.

Otra característica que estudiamos es el dominio de primer nivel de residencia de los sitios de cada componente. El resultado se muestra en la Tabla 11.

Componente	Total Sitios	cl	com	org	net	otro
MAIN IN	1.99 %	99.09 %	0.75 %	0.05 %	0.11 %	0 %
MAIN OUT	3.92 %	99.37 %	0.38 %	0.11 %	0.14 %	0 %
MAIN MAIN	3.76 %	98.33 %	1.39 %	0.17 %	0.11 %	0 %
MAIN NORM	4.36 %	99.68 %	0.22 %	0.07 %	0.02 %	0 %
IN	7.98 %	99.79 %	0.17 %	0.03 %	0 %	0.01 %
OUT	21.52 %	99.17 %	0.66 %	0.08 %	0.09 %	0 %
TOUT	3.81 %	99.55 %	0.42 %	0.03 %	0 %	0 %
TIN	2.9 %	99.67 %	0.29 %	0.04 %	0 %	0 %
TUNNEL	0.27 %	99.6 %	0.4 %	0 %	0 %	0 %
ISLAND	49.49 %	99.73 %	0.19 %	0.04 %	0.04 %	0 %

Cuadro 11: Distribución de los dominios en que están los sitios de cada componente.

4. Características de los Dominios

Definimos el dominio de una página como un sufijo de su nombre de sitio Web, siguiendo la siguiente regla: si el sitio es de la forma `www.A.cl` o `www.xxx.A.cl`, entonces el dominio es `A.cl`. En total se encontraron 158.853 dominios distintos.

4.1. Dirección IP y proveedor de hosting

Realizamos una búsqueda DNS de la dirección IP de cada uno de los sitios estudiados, pudiendo contactar en ese momento al 66 % de ellos. Los sitios que no se pudieron contactar con gran probabilidad ya no existen.

Agrupamos las direcciones IP por dominio, de manera de contar para cuántos dominios distintos se utilizaba la misma IP. La distribución del número de dominios por IP se muestra en la Figura 30.

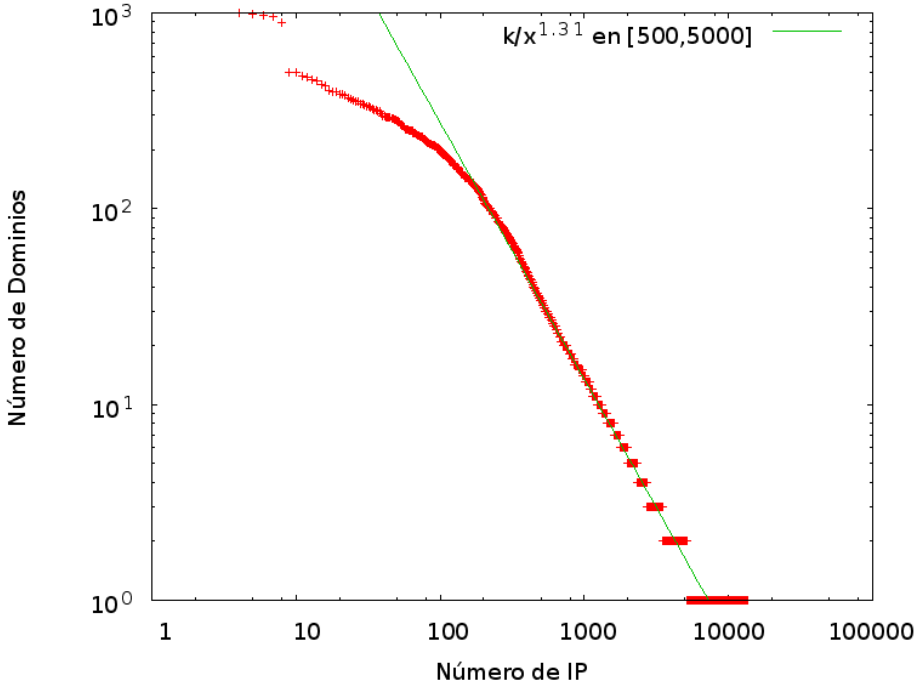


Figura 30: Distribución del número de dominios por dirección IP.

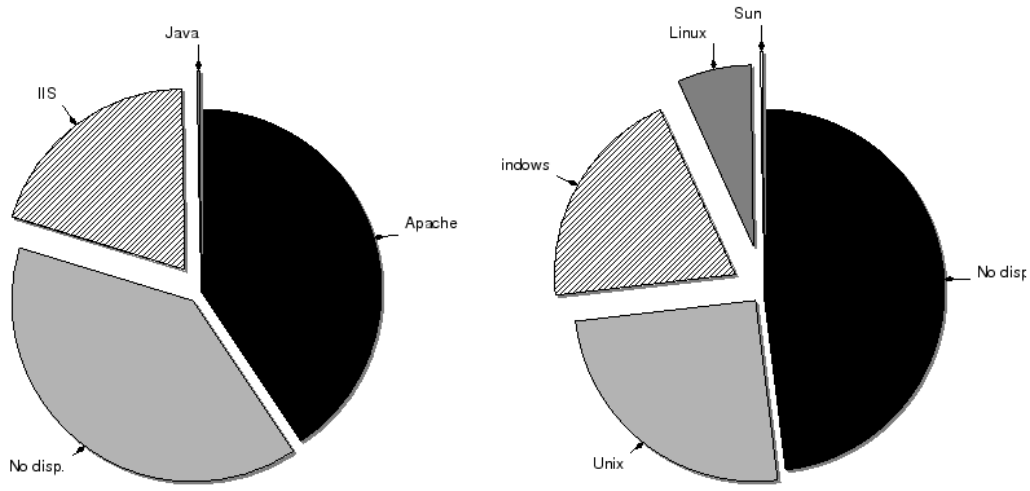
En total hay cerca de 13.500 direcciones IP para todos los dominios. Esto significa que cada dirección tiene en promedio 12 sitios; sin embargo, la distribución es muy sesgada; existen 3 direcciones IP con más de 1.000 dominios distintos cada una, y 8.391 direcciones IP que sólo hospedan 1 dominio. El parámetro de ajuste de la ley de Zipf es 1,31.

4.2. Software utilizado como servidor

Para cada dirección IP examinamos cuál es el software para servidor Web que se utiliza y cuál sistema operativo. Esto se realiza mediante un requerimiento `HTTP HEAD` que solicita solamente el encabezado de la página inicial de un sitio. Una respuesta típica tiene la siguiente forma:

```
HTTP/1.1 200 OK
Server: Apache/1.3.33 (Debian GNU/Linux) PHP/4.3.10-9 mod_ssl/2.8...
```

En algunos casos –como en el ejemplo– la información es bastante completa, incluyendo el nombre del servidor (Apache), la versión (1.3.33), el sistema operativo (Linux) y las extensiones instaladas (PHP y ModSSL). La distribución del tipo de servidor y de sistema operativo se muestra en la Figura 38.



(a) Servidores. (b) Sistemas Operativos.

Figura 31: Distribución del tipo de servidor Web y el sistema operativo por dirección IP.

Los dos servidores dominantes son Apache y Microsoft IIS (Internet Information Server), con Apache doblando la participación de IIS. Sin embargo, un 39 % de las IP's no entrega información sobre el servidor Web que está utilizando, por lo que no se puede decir con certeza que Apache tiene la mayor participación en el mercado. Conociendo la tendencia histórica de la Web chilena probablemente una parte considerable de los servidores no determinados sea Apache. En la Web mundial la proporción de servidores es 69 % para Apache y 21 % para IIS [33].

El mismo fenómeno ocurre con la información de sistema operativo: Unix/Linux tiene un 31 % de participación versus un 20 % de Windows, aunque en un 48 % de los casos no se entrega información sobre el sistema operativo, por lo que no se puede determinar claramente cuál tiene mayor presencia. Si los casos no determinados se distribuyeran proporcionalmente de acuerdo a los casos conocidos, Windows habría disminuido la brecha con los sistemas operativos de código abierto. No es así en todos los países: en España se invierten los papeles, la presencia de Windows alcanza el 43 %, y la de Unix/Linux es de 41 % [9].

4.3. Número de sitios por dominio

En promedio encontramos 1,08 sitios por dominio. Existen 155.784 dominios con solamente un sitio, aunque varios dominios superan con creces el promedio. La distribución del número de sitios para cada uno de los 10.000 dominios más grandes se muestra en la Figura 32.

Los 30 dominios con más sitios que encontramos se muestran en la Tabla 12. Algunos dominios fueron imposibles de contactar nuevamente, y se marcan como "Imposible conectar".

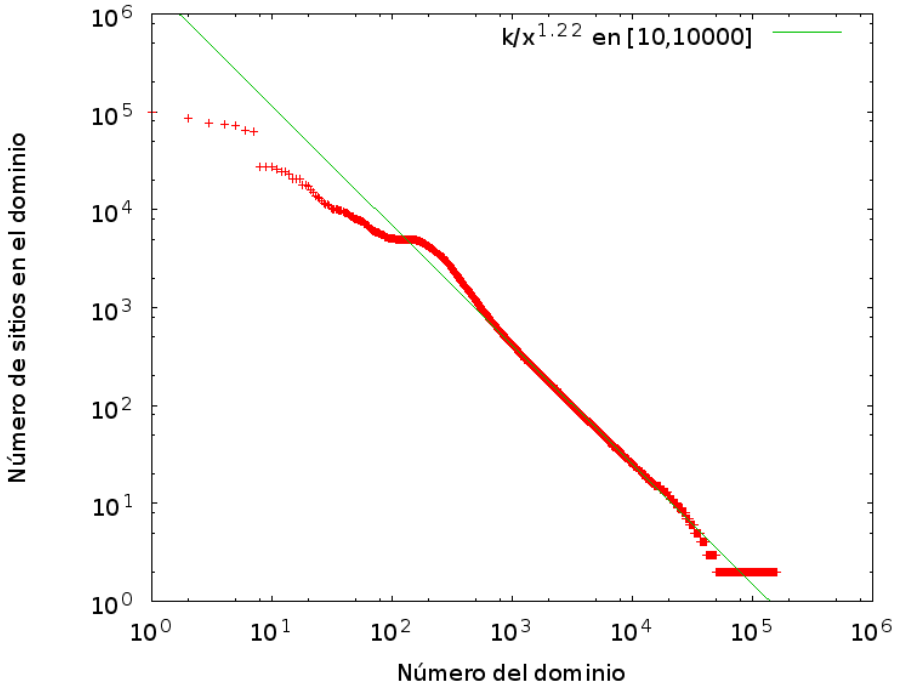


Figura 32: Distribución del número de sitios por dominio.

Sitios	Páginas	Dominio	Sitios	Páginas	Dominio
1407	3641	portalciudadano	104	13125	usach
499	86174	uchile	92	6455	utalca
459	17052	boonic	90	14495	gob
340	2571	scd	84	26419	canal13
329	74542	terra	73	16493	uach
308	10083	ucn	67	6953	123
285	117	notarial	59	1097	zp
249	3331	co	59	20370	ubiobio
156	23083	gov	54	5467	ufro
146	25019	utfsm	51	2361	olx
142	20639	puc	50	1970	contador
132	969	tie	48	9086	uandes
128	16525	ucv	40	10325	udp
127	17013	vivastreet	39	1002	dm
111	20335	udec	39	1277	ulagos

Cuadro 12: Dominios con mayor número de sitios.

4.4. Número de páginas por dominio

Hay un promedio de 47 páginas por dominio. La distribución del número de páginas por dominio es muy sesgada, y sigue una ley de potencias con parámetro 1,67 en su parte central, comparable con el parámetro 1,18 en la Web de España [9]. La Figura 33 muestra la distribución de esta variable, muy similar a la de páginas por sitio que se muestra en la Figura 20.

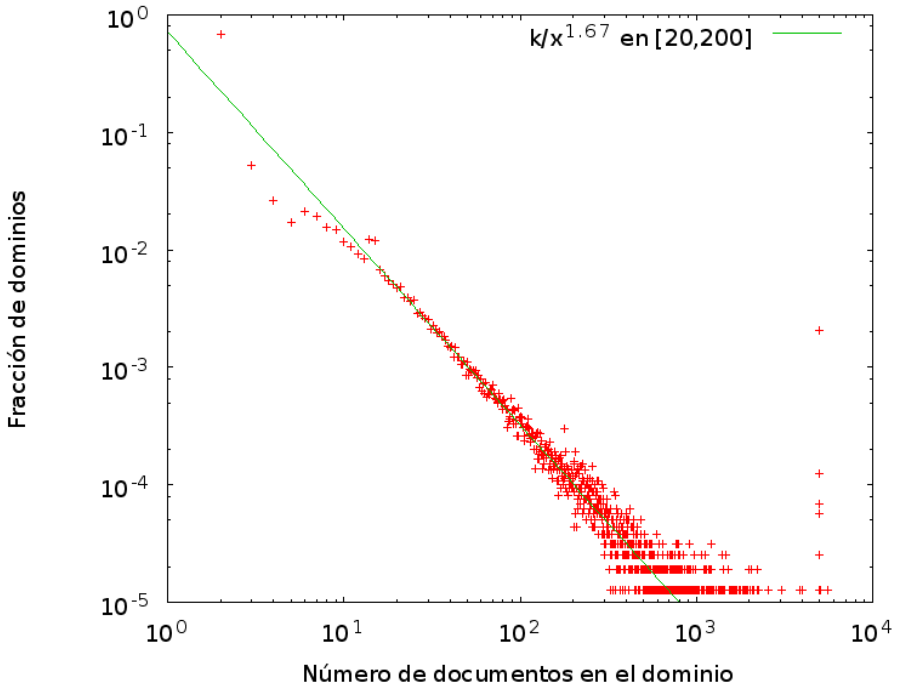


Figura 33: Distribución del número de páginas por dominio.

Hay 34.810 dominios con una única página Web, lo que representa un 21 % de los dominios, prácticamente la misma proporción respecto a los sitios (21,4 %). Este comportamiento es totalmente distinto al que presenta la Web Española, donde la proporción de dominios de una sola página es mucho menor.

4.5. Tamaño total de los dominios

El tamaño promedio de un dominio Web completo es de aproximadamente 328 KiB. La distribución del tamaño total de páginas por dominios se muestra en la Figura 34, y sigue una ley de potencias con parámetro 1,28.

Los dominios con mayor cantidad de texto se muestran en la Tabla 13. Una gran mayoría corresponde a sitios de cadenas comerciales o de remates, Este comportamiento es similar al de la Web de Corea del Sur [10], mientras que en España [9] y Tailandia [37] los primeros lugares corresponden a sitios de instituciones educacionales o del gobierno.

4.6. Títulos de las páginas en un dominio

A continuación analizamos la distribución de títulos distintos por dominio. Medimos la razón entre títulos distintos y páginas de un sitio; por ejemplo, si un sitio tiene 10 páginas y 4 títulos distintos, entonces el valor de este parámetro es 0,25. Esta distribución se aprecia en la Figura 35.

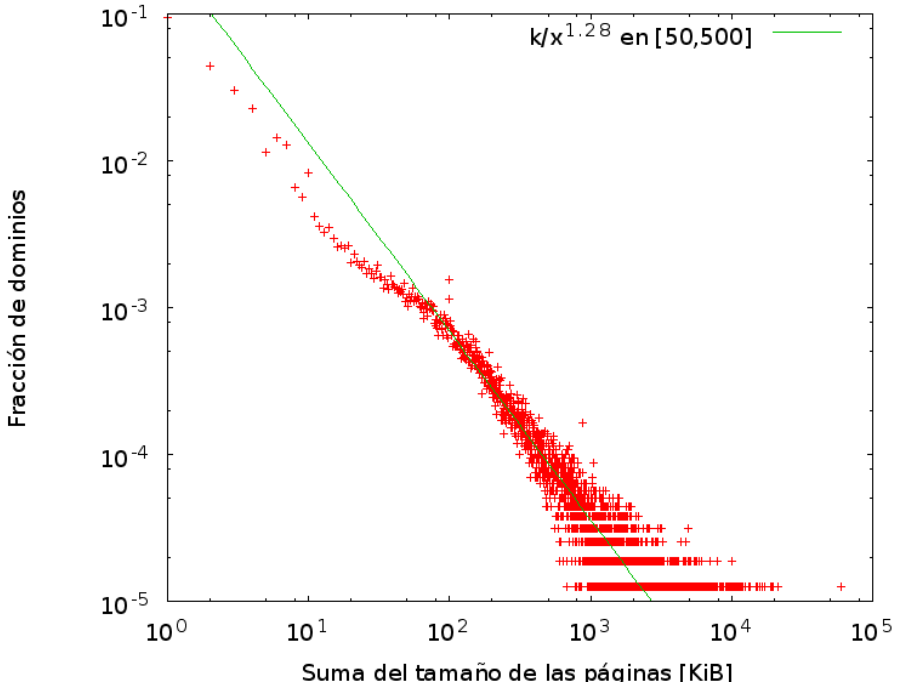


Figura 34: Tamaño total por dominio, considerando sólo el texto.

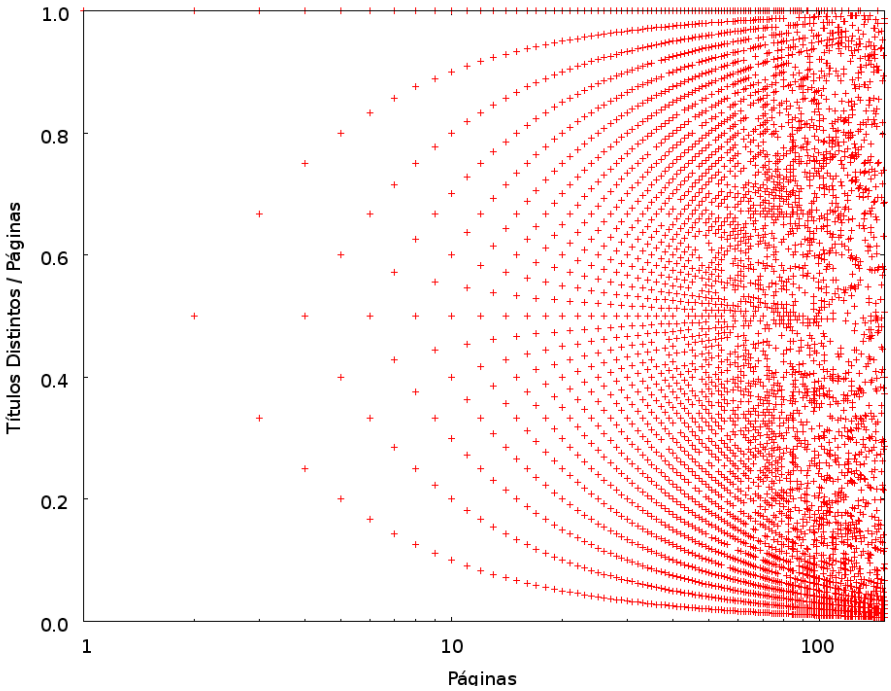


Figura 35: Distribución del número de títulos distintos versus el número de páginas de cada dominio.

Tamaño[MiB]	Nombre del Dominio	Tipo
2131	decompras	Comercial
1782	buy7	Comercial
1313	qsale	Comercial
1173	uchile	Educacional
1107	terra	Comercial
900	k21	Comercial
648	deremate	Comercial
603	mercadolibre	Comercial
569	canal13	Comercial
480	laguiachile	Comercial
457	gov	Gobierno
454	rave	Comercial
450	latercera	Comercial
425	rvt	Comercial
418	almacenesparis	Comercial
408	tercera	Comercial
405	boonic	Comercial
405	vivastreet	Comercial
401	lanaciondomingo	Comercial
394	lnd	Comercial
388	almacenes-paris	Comercial
386	diariolanacion	Comercial
378	fo	Comercial
372	gob	Gobierno
370	bookings	Comercial
369	booking	Comercial
354	lanacion	Comercial
348	concilio	Comercial
341	atinachile	Comercial
338	hispavista	Comercial
330	futurix	Comercial
330	uach	Educacional
329	kontent	Comercial
321	vmf	Comercial
310	cooperativa	Comercial

Cuadro 13: Dominios con mayor cantidad de texto.

En general no observamos una correlación significativa entre estas dos variables: un sitio grande puede tener la misma proporción de títulos distintos que uno pequeño, por lo que este parámetro depende más de la calidad de diseño y planteamiento de un sitio Web que de su magnitud. Sin embargo, la densidad es mayor en la parte inferior del gráfico, lo que indica que es algo más difícil en los dominios grandes tener varios títulos distintos para sus páginas.

4.7. Enlaces entre dominios

A continuación medimos el número de enlaces entre dominios, con el propósito de obtener una representación gráfica de las relaciones entre ellos. En la Figura 36 hemos incluído los 50 dominios que más se relacionan mediante enlaces en la Web Chilena. Los dominios se dividen en tres grupos: comercial (rectángulos), educacional (elipses) y gobierno (rombos), y una línea más oscura significa un mayor número de enlaces.

Utilizamos el programa `neato` del parquete `graphviz` [25], que mediante un modelo de resortes y un algoritmo iterativo encuentra una configuración de baja energía para el grafo. La entrada al programa es un largo mínimo para cada arco, en nuestro caso inversamente proporcional al número de enlaces, y el número de enlaces de cada arco.

El dominio de los sitios comerciales ha desplazado la presencia de sitios de gobierno y educacionales, por lo que es virtualmente imposible apreciar la agrupación entre dominios del mismo tipo que se observó el año 2004 [5] y que se observa en la Web Española [9]; no obstante, algunos sitios, especialmente universitarios, han logrado permanecer unidos.

En la Tabla 14 listamos los 30 dominios que tienen mayor cantidad de enlaces desde otros dominios. Cuando hablamos de dominios y no de sitios sí aparecen dominios educacionales y de gobierno, haciendo más uniforme la distribución, a pesar de que el primer lugar lo tenga un sitio comercial. También aparecen medios de comunicación y sitios gubernamentales de ciencia y cultura.

4.8. Dominios de primer nivel

Nuestra colección de páginas incluye servidores que están en territorio chileno pero que no necesariamente corresponden al dominio .cl. La Tabla 15 muestra como se distribuyen los dominios de primer nivel de la Web de Chile.

Se aprecia que en Chile el dominio más valorado es el dominio de primer nivel nacional, con casi la totalidad de los sitios. Es probable que existan otros sitios con dominios genéricos o extranjeros que se encuentren en la red chilena pero de los cuales no se tiene conocimiento pues no tienen un dominio `.cl` adicional o no reciben enlaces desde otros sitios de la Web nacional.

4.9. Dominios externos de primer nivel

Encontramos más de 90.000.000 de enlaces hacia páginas o documentos fuera de Chile. Los 30 dominios más referenciados se muestran en la Tabla 16, donde se indica el ranking dentro de la Web chilena, el ranking de uso de ese dominio en la Web global [1], el nombre del dominio y el porcentaje de los enlaces que le corresponde dentro del total.

La Figura 37 muestra la distribución de los enlaces a estos dominios. Se observa una ley de potencias con parámetro 4,15 en su parte central, aunque los dominios más importantes en términos de enlaces se alejan de la estimación. Esta distribución está mucho más sesgada que en la Web Española, que tiene parámetro 1,80 y donde el dominio `.com` recibe solamente un 50 % de los

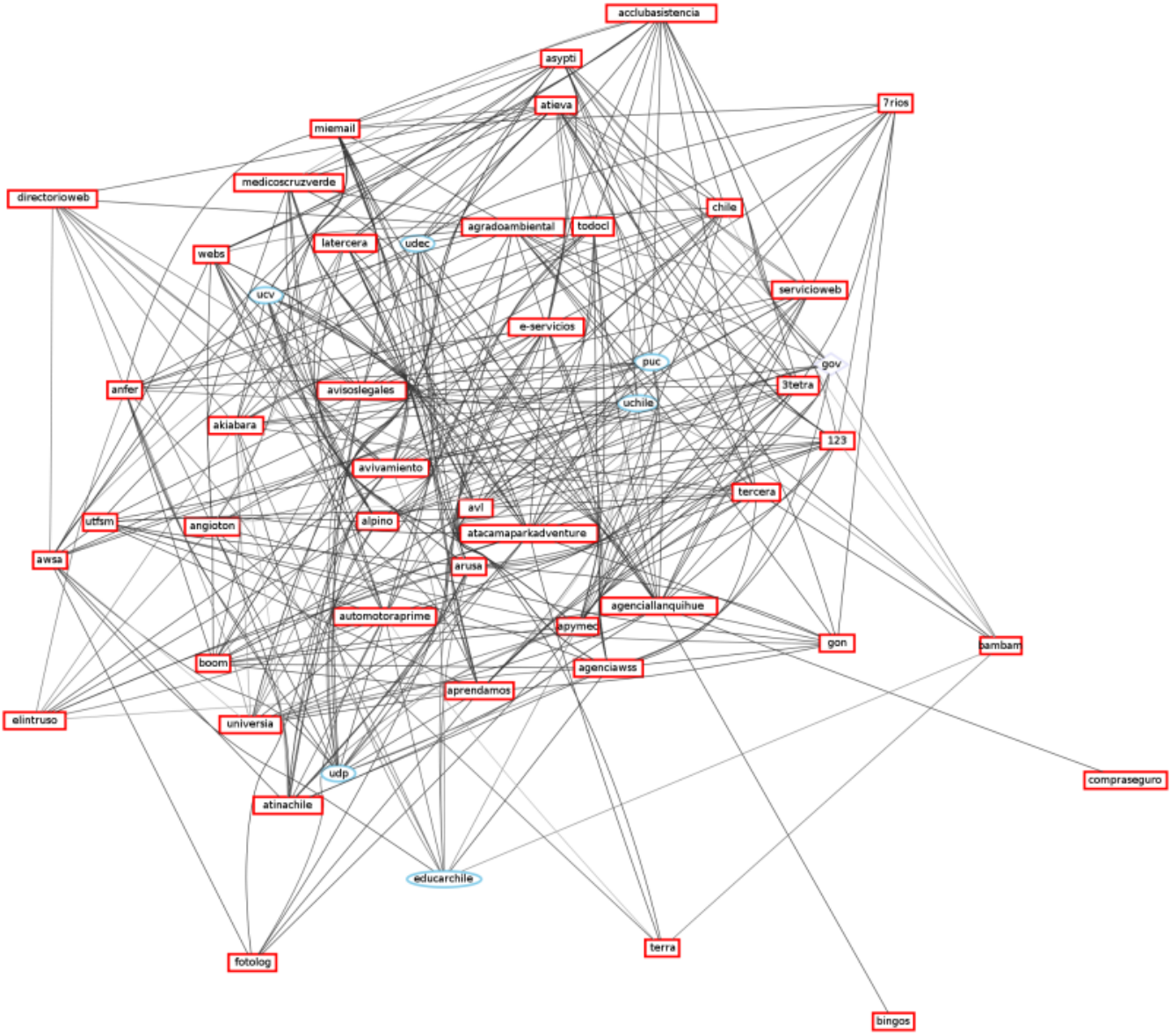


Figura 36: Representación gráfica de los enlaces entre dominios.

Enlaces	Nombre del Dominio	Tipo
24610	boonic	Comercial
6016	uchile	Educacional
4648	vivastreet	Comercial
2402	olx	Comercial
1773	puc	Educacional
1496	terra	Comercial
1450	portalciudadano	Comercial
1307	123	Comercial
1197	sii	Gobierno
1171	gov	Gobierno
1158	udec	Educacional
1083	utfsm	Educacional
1065	gob	Gobierno
1059	mineduc	Gobierno
1015	scd	Comercial
918	latercera	Comercial
906	bcentral	Gobierno
804	meteochile	Gobierno / Cultura, Educación o Ciencia
801	ucv	Educacional
742	canal13	Comercial
718	utalca	Educacional
645	usach	Educacional
635	corfo	Gobierno
634	uach	Educacional
600	sernatur	Gobierno
577	conicyt	Gobierno / Cultura, Educación o Ciencia
575	conama	Gobierno
572	tercera	Comercial
570	sence	Gobierno
567	rvt	Comercial
561	gobiernodechile	Gobierno
512	tvn	Comercial
497	lanacion	Comercial
491	minsal	Gobierno
482	ufro	Comentario

Cuadro 14: Dominios con mayor cantidad de enlaces desde otros dominios.

Dominio	Nombre	% sitios	% páginas
cl	Chile	99.62 %	98.12 %
com	Comercial (Genérico)	0.29 %	1.59 %
net	Red (Genérico)	0.04 %	0.24 %
org	Organización (Genérico)	0.04 %	0.05 %

Cuadro 15: Dominios más usados en la Web de Chile.

enlaces [9]. Nótese que el gráfico continúa más allá de los 200 o 300 dominios válidos por la presencia de errores tipográficos en los nombres de dominio, como `.con` y `.orq`.

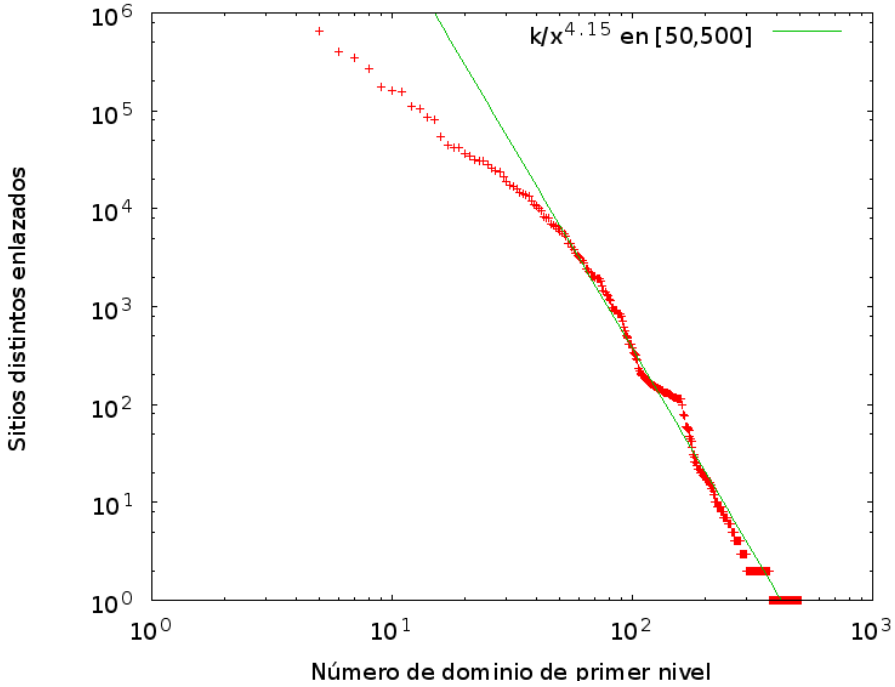


Figura 37: Frecuencia de enlaces a sitios distintos externos, agrupados por dominio de primer nivel.

Los enlaces tienen destinos heterogéneos, al contrario de lo que podría pensarse, pues sólo 8 países de la lista son de habla hispana. Algunos dominios tienen una posición similar a la del ranking de uso de estos dominios en la Web global, como `.com`, `.net`, `.de`, `.br` y `.ec` [1]. El resto de los dominios presenta diferencias en su ranking, siendo la más considerable la del dominio de las Islas Tokelau (`.tk`). [7]

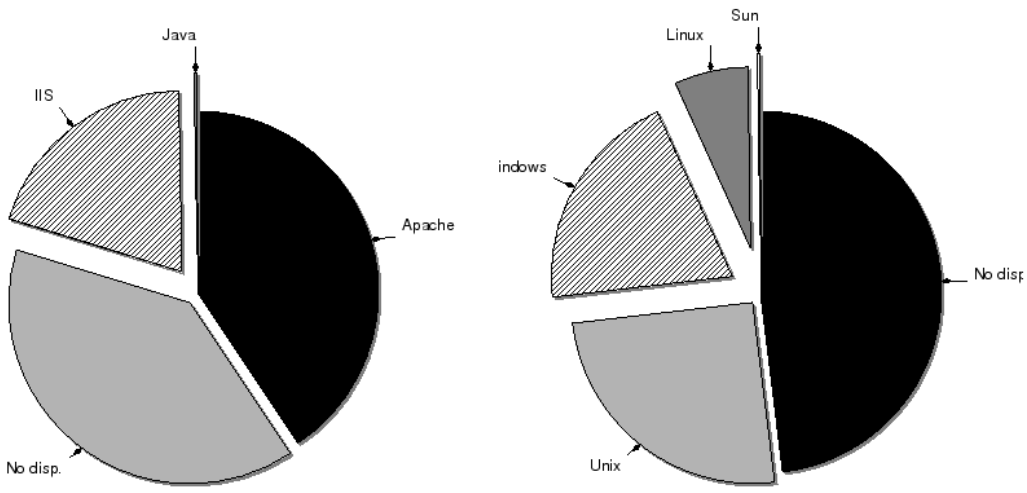
Utilizamos datos de intercambio comercial con el exterior de Aduana de Chile [23], y los comparamos con el número de enlaces encontrados. Los resultados se muestran en la Figura 39(a) para las importaciones y en la Figura 39(b) para las exportaciones. Hay una relación significativa entre el número de enlaces y el volumen de intercambio comercial, y las desviaciones más significativas de esta regla se aprecian en los países asiáticos, que posiblemente debido a una barrera de lenguaje

[7] Este dominio tiene cierta popularidad en Chile ya que muchos chilenos tienen sitios personales en servidores gratuitos (que no están hospedados en Chile) a los cuales les inscribe gratuitamente un dominio `.tk` a cambio de publicidad en la página de inicio de sus sitios.

Ranking	Rank. Global	Dominio	Nombre	Sitios
1	2	com	Comercial (genérico)	61,42 %
2	28	org	Organización sin fines de lucro (genérico)	13,23 %
3	1	net	Red (genérico)	7,08 %
4	29	ar	Argentina	3,62 %
5	71	info	Información (genérico)	2,25 %
6	20	es	España	1,91 %
7	5	de	Alemania	1,44 %
8	77	biz	Negocios (genérico)	0,98 %
9	12	uk	Reino Unido	0,88 %
10	16	mx	México	0,87 %
11	21	us	Estados Unidos	0,61 %
12	11	br	Brasil	0,55 %
13	188	tk	Tokelau	0,48 %
14	6	edu	Educacional (genérico)	0,43 %
15	4	it	Italia	0,29 %
16	7	fr	Francia	0,25 %
17	8	nl	Países Bajos	0,23 %
18	75	ve	Venezuela	0,23 %
19	18	be	Bélgica	0,20 %
20	41	gov	Gobierno EE.UU.	0,18 %
21	52	pe	Perú	0,18 %
22	13	pl	Polonia	0,17 %
23	3	jp	Japón	0,17 %
24	63	uy	Uruguay	0,14 %
25	45	co	Colombia	0,13 %
26	89	ec	Ecuador	0,13 %
27	22	ch	Suiza	0,12 %
28	50	ie	Irlanda	0,10 %
29	108	ws	Samoa	0,10 %
30	25	at	Austria	0,09 %

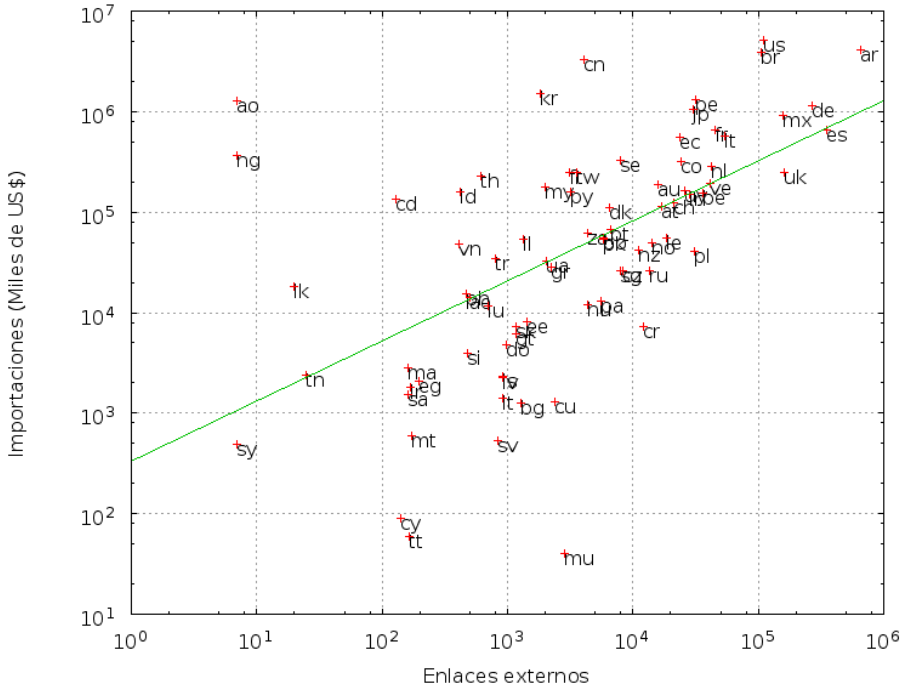
Cuadro 16: Fracción de enlaces a los 30 dominios externos más referenciados.

están más conectados con nosotros en términos de intercambio comercial que en la Web. Esta relación se ha estudiado con mayor profundidad en [6].

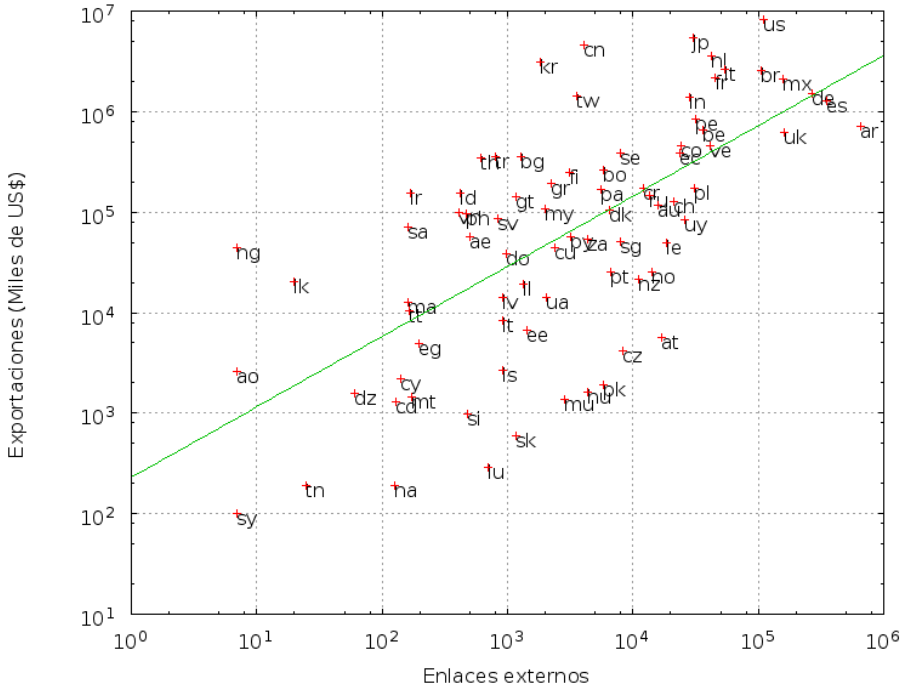


(a) Servidores. (b) Sistemas Operativos.

Figura 38: Distribución del tipo de servidor Web y el sistema operativo por dirección IP.



(a) Enlaces e Importaciones. Correlación: 0,56.



(b) Enlaces y Exportaciones. Correlación: 0,23.

Figura 39: Relación entre el número de enlaces externos y el intercambio comercial de Chile.

5. Conclusiones

Al realizar este Estudio hemos tomado una fotografía de la Web Chilena durante el mes de Agosto de 2006. Esto es similar a lo que hace un astrónomo cuando observa las estrellas en el Universo [2]: lo que ve es la luz que viajó desde las estrellas, que en ese momento ya pudieron dejar de existir, del mismo modo en que ya no existen algunos de los sitios que se mostraron relevantes ante algún indicador, como el tamaño en documentos o en bytes de los sitios.

Se puede decir que las evoluciones más notables de la Web chilena se refieren a su número de páginas, que se duplicó en un período de dos años desde el 2004; o al número de sitios, que se triplicó. Esta explosión ha incidido en la calidad de los sitios: el bajo porcentaje de páginas de calidad, en términos de enlaces, hace dos años se repartía en más de la mitad de los sitios, mientras que el 2006 se repartió sólo en poco más de un tercio de ellos.

De todas las páginas existentes en la Web chilena, un 25 % de ellas fue creada o modificada durante el período 2005–2006. A pesar de ello, es necesario considerar que la mayoría de los usuarios no va muy profundo dentro de los sitios Web; esto significa que hay miles o millones de páginas que son visitadas muy raras vez o casi nunca. De hecho existe una fracción no despreciable de páginas que no han sido modificadas en los últimos 8 años.

No deja de sorprender lo arraigado que está el dominio nacional y la aparente ausencia de spam en la Web chilena; sí tienen presencia sitios que replican contenido entre sí y generan automáticamente miles de páginas con el mismo contenido. A pesar de ello la presencia de sitios de universidades y otras entidades educacionales, como de sitios del gobierno y de medios de comunicación siguen teniendo una presencia importante en términos cuantitativos y cualitativos: los sitios más referenciados prácticamente se repiten.

Un estudio de la Web como el que hemos presentado tiene varias y distintas aplicaciones. La más directa tiene que ver con el desarrollo de mejores sistemas de búsqueda y de estructuras de datos especiales para la Web. Por ejemplo, la presencia de los nuevos CMS más enfocados a los usuarios han puesto algunas barreras a la labor de los recolectores e indizadores: tales sistemas buscan no sólo brindar una mejor experiencia a los usuarios, sino que mejorar sus posiciones en los resultados de búsqueda (fomentando el uso semánticamente correcto de las etiquetas HTML y re-escribiendo las URLs con las palabras clave y títulos de los documentos), pero entorpecen el recorrido de la Web ya que algunas de esas prácticas sesgan las estadísticas.

Otro fin de este informe es mostrar la heterogeneidad de la Web, por un lado positiva debido a su diversidad; por otro, negativa, debido a su poca calidad, por la presencia de numerosos sitios aislados, con poco contenido y pocas referencias.

Es posible observar que esta muestra presenta propiedades estadísticas muy similares a los de otras muestras, lo que indica que puede ser usada en estudios que sean al menos parcialmente extrapolables a la realidad de la red global.

Agradecimientos

Rodrigo Scheihing realizó la operación del recolector durante el proceso de descarga de páginas.

A. Glosario

El siguiente glosario incluye términos básicos de Internet en general y de la Web en particular que son usados en este documento:

AJAX Asynchronous Javascript and XML. Tecnología que permite seguir interactuando al navegador web con el servidor después de cargada una página. Por ejemplo: se utiliza para no tener que recargar una página completa ante una acción puntual del usuario.

CMS Content Management System, Sistema de Administración de Contenidos. Aplicación Web que toma control del manejo y publicación del contenido de un sitio. Por ejemplo: blogs, foros, galerías y aplicaciones personalizadas.

Dominio La forma de asignar nombres de computador en Internet tiene una estructura jerárquica. Un grupo de computadores cuyos nombres comparten un sufijo en común, por ejemplo: "`.cl`" o "`gencat.cl`" constituyen un dominio.

Dirección IP Una secuencia de cuatro números (en el estándar IP versión 4) que identifican la ubicación de cada computador conectado a Internet.

Internet Red internacional que conecta miles de redes más pequeñas. "Internet" con mayúscula se refiere a la red que actualmente se usa, mientras que "internet" con minúscula es el concepto de interconectar varias redes.

Metadatos Datos acerca de una página Web que no son su contenido principal (o "datos acerca de los datos"). Usualmente incluyen su dirección, fecha, tamaño, palabras clave, descripción, etc.

Nombre de Computador (Hostname) Nombre que se asocia a una dirección IP (ejemplo: "`www.cwr.cl`" o "`anakena.dcc.uchile.cl`").

Página Toda entidad en la Web que tiene asociada una URL. En este documento usamos una definición un poco más restrictiva que no considera como páginas a imágenes, vídeo, música y otros archivos multimedia o comprimidos.

Página estática Toda página que existe previamente a ser solicitada.

Página dinámica Toda página que es creada en el momento en que es solicitada.

Servicio Es un programa que se puede ejecutar utilizando Internet. Ejemplos: correo electrónico, chat en línea, World Wide Web.

Servidor Un computador que está conectado a Internet y presta algún servicio.

Sitio Web Nombre de un ordenador que presta el servicio de proveer páginas Web.

TagCloud Nube de Etiquetas. Las etiquetas clasifican material de forma natural y no tan rígida como una taxonomía (Secciones o Categorías). Permiten representar la importancia de algunos términos dentro de una colección de datos de manera gráfica: los términos más importantes tienen mayor tamaño.

URL Estándar para referirse a una dirección en la Web, ejemplo: "`http://www.sitio.cl/pagina.html`". Definido en [LMM94].

World Wide Web También llamada simplemente Web es uno de los servicios que pueden prestar los servidores conectados a Internet.

Referencias

[1] Internet Domain Survey, 2006. `http://www.isc.org/ds/`.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*. Addison-Wesley Harlow, England, 1999.

[3] Ricardo Baeza-Yates and Carlos Castillo. Caracterizando la web chilena. In *Encuentro chileno de ciencias de la computación*, Punta Arenas, Chile, 2000. Sociedad Chilena de Ciencias de la Computación.

[4] Ricardo Baeza-Yates and Carlos Castillo. Relating web characteristics with link based web page ranking. In *Proceedings of String Processing and Information Retrieval SPIRE*, pages 21–32, Laguna San Rafael, Chile, 2001. IEEE CS Press.

[5] Ricardo Baeza-Yates and Carlos Castillo. Características de la web chilena 2004. Technical report, Center for Web Research, University of Chile, 2005.

[6] Ricardo Baeza-Yates and Carlos Castillo. Relationship between web links and trade. *Proceedings of the 15th international conference on World Wide Web*, pages 927–928, 2006.

[7] Ricardo Baeza-Yates and Carlos Castillo. WIRE: Web Information Retrieval Environment, 2006. `http://www.cwr.cl/projects/WIRE/`.

[8] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis Efthimiadis. Characterization of national web domains. *To appear in ACM TOIT*, 2006.

[9] Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. Características de la web de españa. *El Profesional de la Información*, 15(1), January 2006.

[10] Ricardo Baeza-Yates and Felipe Lalanne. Characteristics of the korean web. Technical report, Korea–Chile IT Cooperation Center ITCC, 2004.

[11] Ricardo Baeza-Yates and Bárbara Poblete. Dynamics of the chilean web structure. *Comput. Networks*, 50(10):1464–1473, July 2006.

[12] Ricardo Baeza-Yates, Bárbara Poblete, and Felipe Saint-Jean. Evolución de la web chilena 2001–2002. Technical report, Center for Web Research, University of Chile, 2003.

[13] Albert-László Barabási. *Linked: The New Science of Networks*. Perseus Books Group, May 2002.

[14] A.A. Benczur, K. Csalogany, D. Fogaras, E. Friedman, T. Sarlos, M. Uher, and E. Windhager. Searching a small national domain–a preliminary report. *Poster Proceedings of Conference on World Wide Web*, 2003.

[15] T. Berners-Lee, L. Masinter, and M. McCahill. RFC1738: Uniform Resource Locators (URL). *Internet RFCs*, 1994.

[16] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African web. *The Eleventh International WWW Conference, May*, 2002.

[17] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. *Proceedings of the ninth WWW Conference*, 2000.

[18] J. Cho, N. Shivakumar, and H. Garcia-Molina. Finding replicated web collections. *ACM SIGMOD*, pages 355–366, 1999.

[19] Brian D. Davison. Topical locality in the web. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279, New York, NY, USA, 2000. ACM Press.

[20] S. Dill, R. Kumar, K.S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-Similarity In the Web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[21] Efthimis Efthimiadis and Carlos Castillo. Charting the Greek Web. In *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*, Providence, Rhode Island, USA, November 2004. American Society for Information Science and Technology.

[22] D. Fetterly, M. Manasse, and M.Ñajork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, pages 1–6, 2004.

[23] Servicio Nacional de Aduanas Gobierno de Chile. Estadísticas, 2006.

[24] D. Gomes and M.J. Silva. A characterization of the portuguese web. *3rd ECDL Workshop on Web Archives, Trondheim, Norway*, 21, 2003.

[25] Graphviz. Graph Visualization Software, 2006. `http://www.graphviz.org`.

[26] The PHP Group. PHP: Hypertext Preprocessor, 2006. `http://www.php.net`.

[27] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM Press.

[28] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[29] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[30] Guowei Liu, Yong Yu, Jie Han, and Guirong Xue. China web graph measurements and evolution. In *Web Technologies Research and Development (APWeb)*, pages 668–679, Shanghai, China, 2005. Springer Berlin / Heidelberg.

[31] Microsoft. ASP: Active Server Pages, 2006. `http://msdn.microsoft.com/asp.net/`.

[32] Marco Modesto, Álvaro Pereira, Nivio Ziviani, Carlos Castillo, and Ricardo Baeza-Yates. Um novo retrato da web brasileira. In *Proceedings of XXXII SEMISH*, pages 2005–2017, São Leopoldo, Brazil, 2005.

[33] Netcraft. Netcraft, 2006. `http://www.netcraft.com`.

[34] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[35] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. *8th Annual International Computing and Combinatorics Conference (COCOON)*, pages 330–339, 2002.

[36] A. Rauber, A. Aschenbrenner, O. Witvoet, R.M. Bruckner, and M. Kaiser. Uncovering Information Hidden in Web Archives. *D-Lib Magazine*, 8(12):1082–9873, 2002.

[37] S. Sanguanpong, P.P. Nga, S. Keretho, Y. Poovarawan, and S. Warangrit. Measuring and analysis of the Thai World Wide Web. *Proceeding of the Asia Pacific Advance Network conference*, pages 225–230, 2000.

[38] T. Suel and J. Yuan. Compressing the graph structure of the web. *Data Compression Conference (DCC)*, pages 213–222, 2001.

[39] M. Thelwall and D. Wilkinson. Graph structure in three national academic Webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8):706–712, 2003.

[40] Gabriel H. Tolosa, Fernando R. Bordignon, and Pablo J. Lavallén. Caracterización del espacio web de perú. 2006.

[41] G.H. Tolosa and F.R.A. Bordignon. Análisis de Enlaces en el Espacio Web de las Universidades Argentinas. 2006.

[42] Eveline A. Veloso, Edleno de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender, Ribeiro B. Neto, and Nivio Ziviani. Um retrato da Web Brasileira. In *Proceedings of Simposio Brasileiro de Computacao*, Curitiba, Brasil, 2000.

[43] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.