Evolución de la Web Chilena 2001-2002

Ricardo Baeza-Yates
Barbara J. Poblete
Felipe Saint-Jean

Centro de Investigación de la Web
Depto. de Ciencias de la Computación
Universidad de Chile

Enero 2003

Resumen

Este es el segundo estudio [1] sobre las características de la Web Chilena basado en los datos obtenidos por el buscador de páginas web TodoCL[1], en los años 2001 y 2002. Este estudio, en forma similar al primero, presenta datos estadísticos y comparativos de la evolución de las diferentes componentes de la Web Chilena, analizando sus cifras globales, su topología y las características de las consultas formuladas por los usuarios de TodoCL. Por primera vez la información recopilada hace posible realizar un análisis de la evolución de la Web Chilena entre los años 2000, 2001 y 2002 a nivel de las páginas, sitios y dominios que la componen.

1. Introducción

La Web se caracteriza por estar definida por un conjunto inusual y heterogéneo de elementos. Las mismas características que la hacen un importantísimo medio de difusión y comunicación, hacen muy complejo su análisis. Debido a esto surge el interés y la importancia de realizar periódicamente una descripción de sus principales características y de su evolución en el tiempo. Dado el enorme tamaño de la Web en la actualidad se hace fundamental el estudio de subconjuntos de esta, en el caso de este estudio se analizará la Web Chilena, a través de los datos recopilados por el buscador chileno TodoCL, parte del *spin-off* Barcino Ltda. del Departamento de Ciencias de la Computación de la Universidad de Chile, en colaboración con Akwan (Brasil).

Este estudio recopila los datos recogidos por TodoCL en el periodo del 2001 y 2002, lo cual permite llevar a cabo diversas comparaciones entre estos dos años.

El análisis realizado se divide en tres partes principales. En la primera parte se estudian los contenidos de la Web Chilena, principalmente el número de elementos encontrados a nivel de páginas, sitios y dominios. Destacando que una gran parte de los sitios y dominios chilenos poseen sólo una página, concentrándose de esta forma la mayor parte del contenido en unos pocos sitios. También se presentan estadísticas de los sitios de mayor tamaño en Mbytes de la Web Chilena, y del uso de los diferentes medios y formatos. La segunda parte de este estudio se refiere a la topología de la Web Chilena dado por un análisis de su conectividad a nivel de sitios y dominios. En este análisis se pueden ver las componentes más importantes de la Web Chilena y a su vez se puede observar como han ido variando los sitios y dominios que las conforman en la medida que pasa el tiempo.

[1] Localizado en http://www.todocl.cl

También es importante observar el número de sitios que han ido desapareciendo de la Web en cada una de las componentes. La tercera y última parte de este estudio se enfoca en la información proporcionada por los usuarios de TodoCL, es decir, en las consultas que ellos realizan a través del buscador. Esto permite observar el conjunto de las palabras más buscadas en Chile. Para finalizar se realizaran algunas conclusiones principales del estudio.

2. Conceptos Básicos

2.1. Buscadores

El buscador utilizado para este estudio, TodoCL, es un buscador de indexación automática, al igual que Google[2]y AlltheWeb[3] El proceso de recolección de páginas realizado por estos buscadores tiene dos componentes principales, un *recolector de páginas* que es un programa que comienza recorriendo e indexando sitios predeterminados, estos puntos son los puntos de partida, para luego seguir recorriendo todos aquellos sitios que son apuntados por los primeros en forma recursiva. La otra componente, en la recolección de páginas, es el *planificador*, que se encarga de coordinar el funcionamiento simultáneo de varios recolectores.

Para obtener los datos necesarios para este estudio, se utiliza entonces el recolector y el scheduler de TodoCL. Como puntos de partida TodoCL utiliza principalmente páginas bajo el dominio .CL más algunas páginas en el dominio .NET y .COM pertenecientes a empresas Chilenas. Para el procedimiento recursivo de recolección de páginas TodoCL recorre e indexa todas las páginas Chilenas que encuentra el recolector en su camino.

Cabe destacar que en el proceso de recolección no sólo se indexan páginas HTML, sino que también son indexados el texto de páginas en formato PDF, PostScript y Word, después de ser pasadas por un filtro.

Los archivos binarios (.MP3, .AVI, WAV, etc.) no son recolectados y por lo tanto no se incorporan a la colección.

Otros conceptos importantes en cuanto a los buscadores son:

- **Colección**: Son todos los documentos recolectados e indexados por el buscador.
- **Página**: Un documento indexado por el buscador.

[2]Localizado en http://www.google.com.
[3]Localizado en http://www.alltheweb.com.

- **Sitio**: Es un servidor Web (lógico) identificado por un subdominio, por ejemplo: *dcc.uchile.cl* que es un sitio perteneciente al dominio *uchile.cl*
- **Dominio**: En el caso de Chile, es cualquier nombre de la forma $x.y$ donde $y = .cl$

2.2. Zipf

La ley de Zipf lleva el nombre del profesor de lingüística de Harvard, George Kingsley Zipf (1902-1950). Es básicamente una distribución en la cual, si definimos P_i como la frecuencia de ocurrencia del i-ésimo evento más frecuente, tendremos que

$$P_i \sim \frac{1}{i^a}$$

donde a es una constante cercana a 1, que llamaremos parámetro de la distribución de Zipf. La ley de Zipf es una distribución caracterizada por presentar eventos poco frecuentes y eventos muy frecuentes.

Al ser la ley de Zipf una función exponencial, al graficar P_i en escala logarítmica veremos una línea recta, cuya pendiente será el inverso aditivo del exponente o parámetro de Zipf.

3. El Contenido

3.1. Cifras Globales

En la tabla 1 se muestran los valores de las colecciones de documentos recolectados por el buscador de TodoCL para los años 2000, 2001 y 2002. Estos documentos corresponden a todos los dominios no .CL que se encuentran en Chile y todos los dominios .CL tanto dentro como fuera de Chile, encontrados por el buscador.

En la tabla 1 cabe destacar que no todas las páginas chilenas son recolectadas por el buscador, ya que algunas de ellas están marcadas como páginas no indexables por sus administradores. Por ejemplo, las páginas que aún no tienen dominio DNS asignado muestran una página de cortesía perteneciente al NIC Chile[4] que está marcada como no indexable, de manera de no distorsionar la colección, ya que son muchos los dominios en este estado. De esta misma tabla es importante

[4]Localizado en http://www.nic.cl.

Año	2000	2001	2002
Páginas	730673	794218	2214253
Sitios	10352	21207	39320
Dominios	9102	19389	35520

Tabla 1: Cifras globales de la colección de TodoCL.

Páginas	1988706
Sitios	38307
Dominios	34867

Tabla 2: Cifras globales de documentos .CL en el año 2002.

observar el aumento que se ha producido año a año en las cifras globales de la Web Chilena. Se puede ver que el número de sitios y dominios prácticamente se duplica cada año.

Como se mencionó anteriormente la mayor parte de las páginas que recolecta TodoCL pertenecen al dominio .CL, las cifras exactas de las páginas, sitios y dominios .CL encontrados el año 2002 por el buscador son las de la tabla 2

En la tabla 3 vemos datos sobre las ubicaciones de los sitios chilenos. Este estudio se hizo según en número IP del servidor y su pertenencia o no a redes chilenas. Los números no son del todo consistentes con los anteriores, dado que para muchos sitios el DNS no respondió[5] Es interesante ver la cantidad de sitios del dominio .CL que no están en Chile. No es posible determinar sitios de dominio no .CL chilenos ubicados fuera de Chile.

3.2. Páginas por sitio y dominio

Una porción importante de los dominios inscritos no se utiliza, y de los que se utilizan, muchos tienen sólo una página, la página de presencia. En la Web Chilena, el 56 % de los dominios y el 54 % de los sitios tienen sólo una página. La figura 1 muestra la cantidad de páginas por sitio para el año 2002. En este gráfico se puede observar que la distribución de páginas puede ser aproximada por una *Zipf* con parámetro *1.58* Esta distribución es muy similar a lo observado en el año 2001, representado por una *Zipf* con parámetro *1.8*

[5] Agradecemos a Emilio Davis por su apoyo para obtener estos datos.

CL en Chile	20.457
.CL fuera de Chile	13.334
.COM en Chile	635
.NET en Chile	155
.ORG en Chile	69
.AR en Chile	66
Otros en Chile	88

Tabla 3: Distribución de sitios chilenos para el año 2002.

# Sitios	IP
1000	200.72.1.75
956	209.61.188.26
693	200.54.163.35
611	64.239.33.249
484	216.34.94.186
414	200.24.224.1
392	200.14.114.104
367	200.27.158.7
360	200.54.144.200
331	216.241.9.155
325	200.29.21.60
302	200.29.128.35
237	216.155.73.45
236	200.14.80.128
226	208.185.127.169
222	200.27.135.2
212	200.27.158.10
207	200.29.13.50
195	200.28.216.20
194	216.241.0.130

Tabla 4: IPs de servidores con mayor número de sitios (2002).

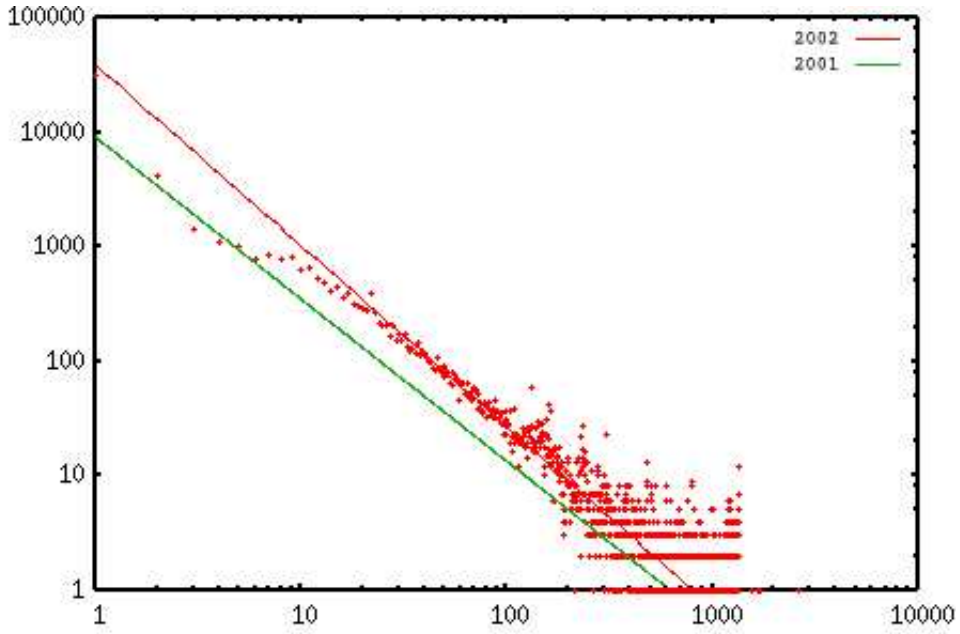

Figura 1: Cantidad de sitios vs. cantidad de páginas.

2001		2002	
# Páginas	# Sitios	# Páginas	# Sitios
1-100	19687	1-100	35047
101-200	686	101-200	1783
201-300	239	201-300	676
301-400	167	301-400	375
401-500	107	401-500	266
501-600	65	501-600	190
601-700	30	601-700	134
701-800	33	701-800	105
801-900	18	801-900	125
901-1000	20	901-1000	81
1001-1100	30	1001-1100	96
1101-1200	64	1101-1200	122
1201-1300	51	1201-1300	175
1301-1400	8	1301-1400	140
1401-1500	2	1401-1500	0
1501-1600	0	1501-1600	1
1601-1700	0	1601-1700	2
1701-1800	0	1701-1800	1
2601-2700	0	2601-2700	1

Tabla 5: Distribución de la cantidad de páginas en los sitios.

En comparación con el 2001, en que un 45 % de los sitios tenía sólo una página, vemos que en el 2002 se produce un aumento porcentual y absoluto en el número de sitios con una sola página. Esto puede explicarse por el gran aumento de dominios inscritos, muchos de los cuales sólo poseen una página de presencia.

En la tabla 6 se pueden ver los sitios con mayor número de páginas en Chile para los años 2001 y 2002.

Al comparar los sitios presentados en ambas tablas es posible observar que todos los 20 sitios con más páginas en el 2001 fueron desplazados al año siguiente. En general los sitios que estaban en la lista de los 20 sitios con más páginas el 2001 aun estan entre los 1000 sitios con más páginas de Chile en el año 2002.

En la tabla 7 se presentan los dominios chilenos con mayor número de sitios para los años 2001 y 2002 respectivamente. Es posible apreciar que, en el caso de los dominios con más sitios, el cambio de un año a otro no es tan radical como en

2001		2002	
Sitio	# Páginas	Sitio	# Páginas
ariadna.puc.cl	1418	www.eclac.cl	2656
bsd.attla.cl	1413	www.c-renta.com	1708
pda.attla.cl	1365	www.sanignacio.cl	1695
deportivo.tercera.cl	1364	www.openbox.cl	1654
www.delcerro.cl	1348	ias.sec.cl	1547
www.ctcreuna.cl	1347	www.losnaranjos.cl	1363
baltazar.conicyt.cl	1327	www.centrolinux.cl	1360
www.ctcinternet.cl	1321	www.fancymusic.cl	1360
www.mercuriovalpo.cl	1312	www.itp.cl	1360
www.diarioaustral.cl	1305	www.mercurioantofagasta.cl	1359
www.labmat.puc.cl	1293	www.nikter.cl	1358
www.fwu.cl	1293	nicolette.nic.cl	1357
www.cuarta.cl	1284	www.agthel.cl	1356
www.terramall.cl	1280	www.chipnews.cl	1356
www.kodak.cl	1271	www.santiagotimes.cl	1356
www.estrellaarica.cl	1268	www.caslab.cl	1355
www.codelco.cl	1268	www.planoinmobiliario.cl	1355
www.australtemuco.cl	1266	foros.ircangol.cl	1355
www.sectormatematica.cl	1265	www.redhat.cl	1354

Tabla 6: Sitios con mayor número de páginas en Chile.

2001		2002	
Dominio	# Sitios	Dominio	# Sitios
co.cl	227	co.cl	228
uchile.cl	159	uchile.cl	222
scd.cl	120	terra.cl	182
terra.cl	102	scd.cl	175
puc.cl	68	tripod.cl	135
utfsm.cl	65	puc.cl	91
udec.cl	62	utfsm.cl	73
corp.cl	59	gov.cl	71
gov.cl	40	udec.cl	66
uach.cl	35	usach.cl	60

Tabla 7: Dominios con mayor número de sitios en Chile.

el caso de los sitios con más páginas. La mayoría de los dominios que en el 2001 tenían el mayor número de sitios están presentes nuevamente en el 2002.

3.3. Tamaño

El tamaño promedio de una página en la Web Chilena es de 11.562 bytes, considerando sólo el texto y tags HTML. Sólo el 4 % de las páginas tiene más de 40kb de texto.

El tamaño de los sitios refleja el nivel de contenido que hay en ellos. En el 2000, el 1 % de los sitios más grandes aportaba con el 60 % del tamaño, en el 2001 aportan con el 40 %, lo que indica que la Web es más equilibrada que antes, en el sentido que, proporcionalmente, son más los sitios que aportan contenido.

En la tabla 8 se pueden ver los sitios con mayor contenido en tamaño en Mbytes para el año 2001, considerando el tamaño del sitio completo, es decir, incluyendo archivos no indexables. De esta tabla cabe destacar que la mayoría de estos sitios corresponden a copias locales o *mirrors* del portal de software Tucows[6] Es interesante observar a la vez los datos recopilados en la tabla 9, que representa el contenido en texto plano, sin incluir archivos binarios ni tags HTML, de los sitios en el año 2002.

En la figura 2 se presenta una comparación entre los datos recopilados el año

[6]Localizado en http://www.tucows.com.

Sitio	Tamaño(Mbytes)
pda.attla.cl	71588
www.embnet.cl	57821
www.linuxberg.cl	57003
tucows.ctcinternet.cl	56137
linuxberg.attla.cl	55857
tucows.rdc.cl	55228
tucows.firstcom.cl	54911
tucows.uplink.cl	48871
tucows.telsur.cl	47729
tucows.attla.cl	47149

Tabla 8: Sitios con mayor contenido en Mbytes, incluyendo binarios (2001).

Sitio	Tamaño(Kbytes)
www.plusvalia.cl	33783
www.anfitrion.cl	14522
www.oim.web.cl	14233
www.camara.cl	14023
c6.li2.uchile.cl	13247
sociales.uchile.cl	12848
www.diariooficial.cl	12726
www.diarioficial.cl	12726
rehue.csociales.uchile.cl	12512
ads2.astro.puc.cl	11658
www.csociales.uchile.cl	11295
www.cristiandad.org	10616
www.chiptravel.cl	10274
bitmed.med.uchile.cl	9391
www.freebsd.cl	9065
www.creces.cl	8719
www.quepasa.cl	8696
lucas.linux.cl	8462
www.inteco.cl	8412
www.conama.cl	8159

Tabla 9: Sitios con mayor contenido en Kbytes, sólo texto (2002).

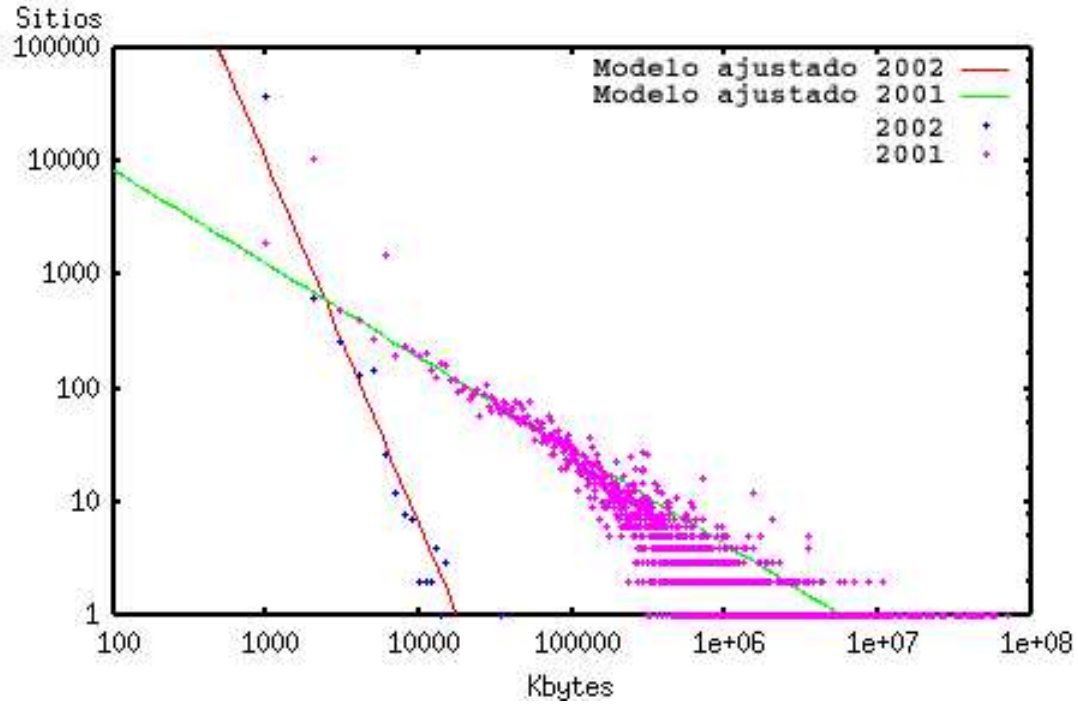

Figura 2: Cantidad de sitios vs. tamaño.

2001, que contienen el tamaño completo de los sitios, y los datos del año 2002, que contienen el tamaño en texto plano de los sitios. A estos datos les fue ajustado una $Zipf$ de coeficientes 0,8 y 3,2, para el 2001 y 2002 respectivamente.

3.4. Medios y formatos

Además del HTML en la Web existen contenidos de diversos tipos, los que también son interesantes de indexar y recuperar. Los documentos de tipo distinto a HTML se separaron en:

Multimedios: Documentos no indexables por el buscador, a su vez se divide en imágenes, video y audio.

Texto: Documentos de texto en formato distinto a HTML, con filtros pueden ser indexados en la mayoría de los casos.

Contenido	Cantidad de Sitios
MAP	36
FLASH	1139
NO-LINKS	6658
PARAM	1162
TOTAL	8995

Tabla 10: Tipos de documentos en sitios con una sola página.

Servidores de aplicación: Son páginas cuyo resultado es HTML, pero son generadas dinámicamente.

Los documentos con los cuales se trabajó fueron seleccionados por el tipo MIME que los describía, no se utilizó ningún algoritmo que detectara su tipo con más seguridad. A continuación se intento determinar la cantidad de documentos distintos encontrados en la Web Chilena. Para esto se determinó la cantidad de archivos con extensiones diferentes, como una aproximación de los tipos de documentos. Cerca de un 85 % del total de documentos (incluyendo multimedios) son HTML o páginas dinámicas que generan HTML. Dentro de los documentos de texto el HTML es un 97 % del total.

En lo que respecta a los documentos multimedios las figuras 3 y 4[7] muestran las distribuciones de los formatos de audio, video, imágenes y documentos de texto que no son HTML, respectivamente.

Respecto a las páginas dinámicas indexadas, la figura 5 permite hacer una comparación de la evolución del uso de ciertos formatos entre el 2001 y el 2002. De estos gráficos se puede apreciar un aumento considerable en el uso de *PHP* entre el año 2001 y 2002 convirtiéndose en el tipo más utilizado en la Web Chilena seguido de cerca por *ASP*

En el año 2001 se hizo un análisis acerca del contenido de los documentos que de los sitios con sólo una página. Los resultados encontrados se muestran en la tabla 10

Los contenidos de estos sitios son los siguientes:

- **MAP**: Son los sitios que tienen mapas de imágenes.
- **FLASH**: Son sitios que tienen “flash”

[7]Datos recopilados el año 2001.

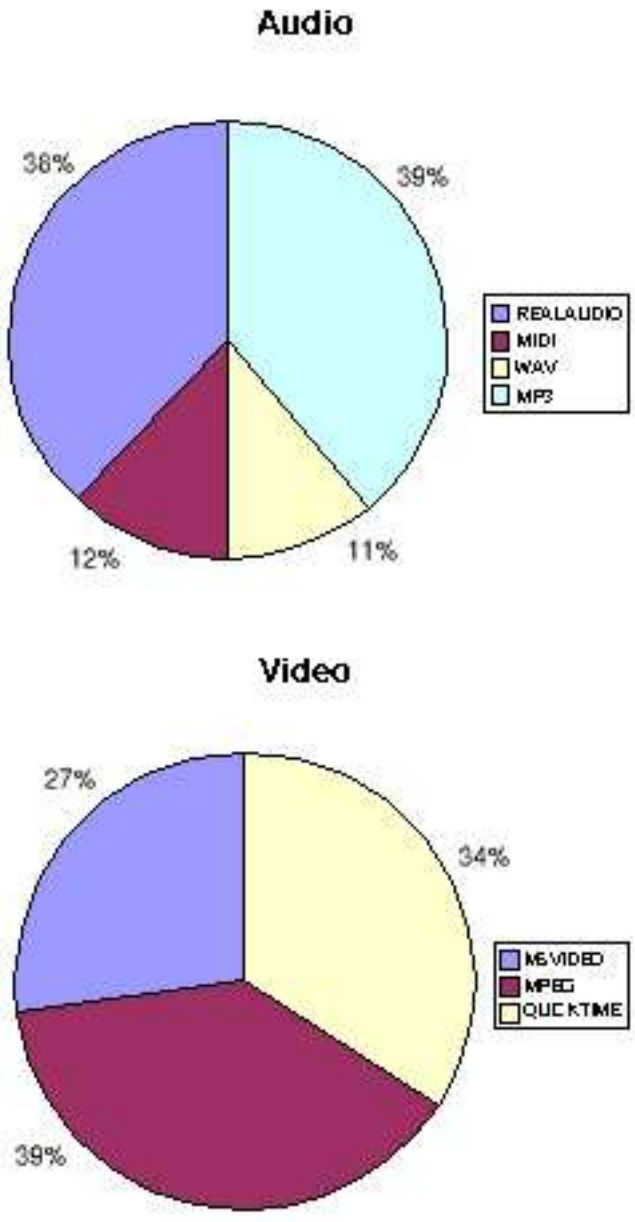

Figura 3: Distribución de documentos de audio y video.

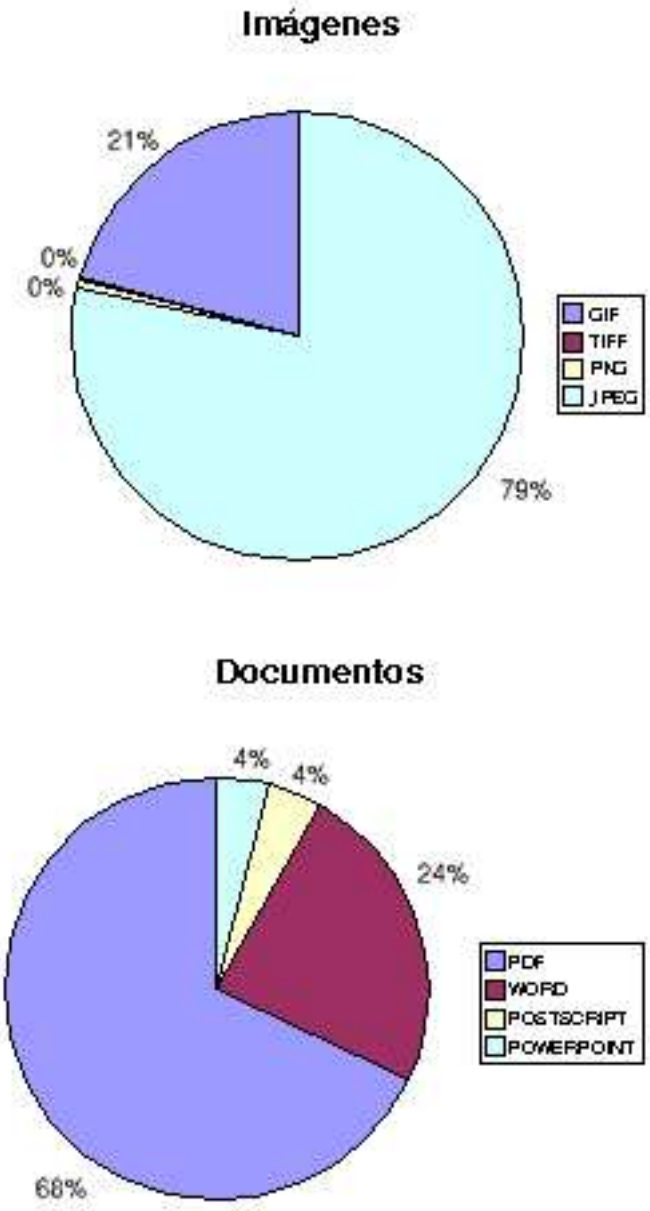

Figura 4: Distribución de documentos de imagen y texto no HTML.

Figura 5: Distribución de páginas dinámicas los años 2001 y 2002.

- **NO-LINKS**: Son los sitios que no tienen links de salida.
- **PARAM**: Sitios que tienen tags de tipo “param”

4. La Topología

En esta sección se describen elementos de la Web Chilena basados en las características topológicas de ésta, es decir, en las páginas, sitios o dominios y la relación de links entre ellos. Al referirnos a links entre sitios (dominios), estamos diciendo que existe al menos un link entre una página de un sitio (dominio) y una página del otro sitio (dominio).

Las características topológicas a nivel de Web son una fuente muy importante de información respecto a ésta. Tanto así que la ubicación topológica de las páginas se ha considerado en Google como la primordial característica de jerarquización, reemplazando a las más clásicas basadas en distancia vectorial entre documento y consulta.

4.1. Enlaces

Los dominios más populares entre administradores de sitios Web son los que se muestran en la tabla 11, la cantidad de referencias mencionada es a nivel de dominio.

La tabla 12 muestra los sitios más referenciados a nivel de sitio. En la Web chilena hay 27.058 dominios fuera del dominio .cl que son referenciados. Los más referenciados son los expuestos en la tabla 13 .

5. Macroestructura

Al analizar la Web como un grafo, es posible aplicar toda la teoría sobre éstos que existe en las matemáticas. Una definición común en la teoría de grafos es la de componente fuertemente conexa; ésta se aplica a un grafo dirigido, donde una componente fuertemente conexa es un subconjunto de los nodos del grafo donde existe un camino entre cualquier par de ellos. Una componente fuertemente conexa en la Web es un conjunto de sitios entre los cuales existen caminos a través de links entre cualquier par de sitios. Las componentes fuertemente conexas en la Web con más de un sitio no son muchas. Llamaremos a la más grande de ellas la componente conexa principal, que además tiende a ser mucho más grande que las

2001		2002	
Dominio	# Referencias	Dominio	# Referencias
uchile.cl	552	uchile.cl	1268
puc.cl	299	sii.cl	415
sii.cl	279	interating.com	408
bcentral.cl	242	puc.cl	407
congreso.cl	229	bcentral.cl	364
elmercurio.cl	228	tercera.cl	337
tercera.cl	214	meteochile.cl	331
conama.cl	205	terra.cl	321
estrategia.cl	204	mineduc.cl	295
eldiario.cl	203	gob.cl	292

Tabla 11: Dominios más referenciados a nivel de dominio.

2001		2002	
Sitio	# Referencias	Sitio	# Referencias
www.uchile.cl	467	contadores.cec.uchile.cl	755
www.puc.cl	329	www.uchile.cl	567
www.sii.cl	306	www.sii.cl	439
www.bcentral.cl	270	m1.interating.com	433
www.congreso.cl	261	www.puc.cl	428
www.elmercurio.cl	255	www.meteochile.cl	399
www.conicyt.cl	255	www.bcentral.cl	392
www.tercera.cl	248	www.mineduc.cl	381
www.udec.cl	243	www.udec.cl	348
www.conama.cl	236	www.tercera.cl	322

Tabla 12: Sitios más referenciados a nivel de sitios.

Dominio	referencias
tucows.com	226495
domaindirect.com	16528
10sites.com	11045
goodyear.com	10123
hispavista.com	9616
philips.com	9424
geocities.com	8673
expowedding.com	7805
kodak.com	7525
microsoft.com	5258
intel.com	4997
cartoonnetwork.com	4883
freeservers.com	4781
oracle.com	4382
themeoftheday.com	4178
yahoo.com	3486
sun.com	3483
hotjobs.com	3003

Tabla 13: Dominios más referenciados fuera de Chile a nivel de páginas (2001).

que le siguen. En [1] se hace la siguiente división de la Web, según su relación con la componente fuertemente conexa principal:

MAIN: Componente fuertemente conexa principal.

IN: Sitios de los que se llega a MAIN, pero de MAIN no se puede llegar a ellos.

OUT: Sitios de los que se puede llegar de MAIN, pero no se puede ir de ellos a MAIN.

TUNNEL: Sitios en caminos de IN a OUT sin pasar por MAIN.

TENTACLE: Sitios a los que se llega de IN o llevan a OUT, pero no están ni en MAIN ni en TUNNEL.

ISLANDS: Sitios no conectados a nada de lo anterior.

Además se utilizaron las extensiones propuestas en [2], que son las siguientes:

MAIN-MAIN: Sitios relacionados directamente con IN y con OUT.

MAIN-IN: Sitios relacionados directamente con IN, pero no con OUT.

MAIN-OUT: Sitios relacionados directamente con OUT, pero no con IN.

MAIN-NORM: Sitios en MAIN que no corresponden a ninguna de las anteriores categorías.

Se decidió hacer el análisis basandose principalmente en sitios, aunque tambien se observará de manera más superficial el comportamiento de los dominios. Lo ideal sería hacer el estudio a nivel de documento, pero un documento está compuesto por varias páginas y no es fácil determinar cuáles son, ya que las relaciones entre éstas son básicamente semánticas. Los sitios, en general, tienen un grupo pequeño de administración y contienen tópicos relacionados. Los dominios, en cambio, pueden agrupar sitios de dedicados a diversos temas.

Al observar la tabla 14 lo que más llama la atención es el gran tamaño de ISLANDS con respecto al resto de las componentes. ISLANDS esta compuesto por cerca de un 50 % de los sitios de la Web Chilena. Los sitios en esta componente tienden a ser siempre los más nuevos, lo que indica que es ahí donde se ha producido gran parte del crecimiento de la Web.

Componente	Tamaño (%) 2000	Tamaño (%) 2001	Tamaño (%) 2002
MAIN	36.45 %	9.25 %	11.98 %
IN	10.79 %	5.84 %	9.97 %
OUT	39.36 %	20.21 %	17.15 %
TUNNEL	0.37 %	0.22 %	0.23 %
TENTACLE-IN	1.32 %	3.04 %	3.11 %
TENTACLE-OUT	4.01 %	1.68 %	3.31 %
ISLANDS	7.68 %	59.73 %	54.21 %
MAIN-MAIN	3.88 %	3.43 %	4.08 %
MAIN-OUT	8.85 %	2.49 %	2.77 %
MAIN-IN	4.76 %	1.16 %	2.24 %
MAIN-NORM	18.95 %	2.15 %	2.88 %

Tabla 14: Comparación del tamaño relativo de las componentes de la Web Chilena.

Año	2000	2001	2002
SITIOS	7497	21207	39320
GONE		1705	5824
NEW		15415	23937

Tabla 15: Cifras generales de las componentes de la Web Chilena a nivel de sitios.

5.1. Composición de las Componentes Actuales

Desde el inicio de la Web los sitios han cambiado su ubicación dentro de ésta. La pregunta que se desea responder es dónde están hoy los sitios de las componentes de hace un año. A continuación se presentan algunas cifras globales de las componentes de la Web Chilena, a nivel de sitios y de dominios, en las tablas 15 y 16 respectivamente.

En la tablas 17 y 18 se analiza el movimiento en las componentes a nivel de sitios y en la tabla 19 se hace el mismo análisis a nivel de dominios. Hay dos lecturas posibles de las tablas 17 y 18 Al ver estas tablas por columnas se puede observar de qué componente vienen los sitios de las componentes actuales. Al estudiarlas por filas vemos dónde están hoy los sitios de las componentes hace un año. La última columna y fila representan los sitios que ya no existen (**GONE**) y los sitios nuevos (**NEW**), respectivamente.

Es interesante notar que OUT y MAIN son componentes altamente estables, ya

Año	2001	2002
DOMINIOS	19389	35520
GONE	-	5266
NEW	-	21397

Tabla 16: Cifras generales de las componentes de la Web Chilena a nivel de dominios.

2000 \ 2001	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	959	724	140	305	11	61	24	509
OUT	195	1151	39	749	5	96	48	668
IN	39	89	118	279	2	31	25	226
ISLANDS	18	124	14	213	0	14	19	174
TUNNEL	1	1	3	18	0	0	2	3
TIN	5	31	0	18	3	3	2	37
TOUT	3	38	25	131	0	4	12	88
NEW	742	2128	901	10955	27	437	225	-

Tabla 17: Composición de las componentes a nivel de sitios en el 2001 respecto del 2000.

que cerca de un 25 % de los sitios que actualmente se encuentran en ellas estaban ahi el año anterior. También se destaca el hecho de que MAIN se compone en un 20 % por sitios que antes estaban en OUT. Sin duda se confirma el hecho que ISLANDS es la componente que más ha crecido y que a la vez es la componente que más sitios ha perdido.

En las figuras 6 y 19 se presenta de forma esquemática como ha sido el movimiento a nivel de sitios y de dominios entre las diferentes componentes de la Web Chilena. El movimiento de los sitios y dominios entre las diferentes componentes, de un año a otro, puede verse reflejado en las partes de las componentes que se representan en un color más claro al original. En estos esquemas se pueden ver reflejados en forma aproximada todos los datos mencionados en las tablas 14, 15, 16, 18 y 19

2001 \ 2002	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	1214	339	158	42	1	17	8	183
OUT	901	1683	188	532	15	128	43	796
IN	233	98	292	196	1	22	16	382
ISLANDS	422	1351	786	5182	23	365	299	4240
TUNNEL	11	15	3	4	1	2	0	12
TIN	78	215	25	128	2	66	5	127
TOUT	52	79	41	59	0	18	24	84
NEW	1801	2965	2430	15173	50	608	910	-

Tabla 18: Composición de las componentes a nivel de sitios en el 2002 respecto del 2001.

2001 \ 2002	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	918	218	79	35	0	4	4	141
OUT	892	1424	167	466	14	97	35	560
IN	206	79	288	182	2	19	9	326
ISLANDS	487	1276	970	4967	25	320	242	4074
TUNNEL	4	1	3	1	0	0	0	4
TIN	88	226	22	134	0	59	8	102
TOUT	35	22	39	35	0	2	19	59
NEW	1376	2176	2644	14171	27	419	584	-

Tabla 19: Composición de las componentes a nivel de dominios en el 2002 respecto del 2001.

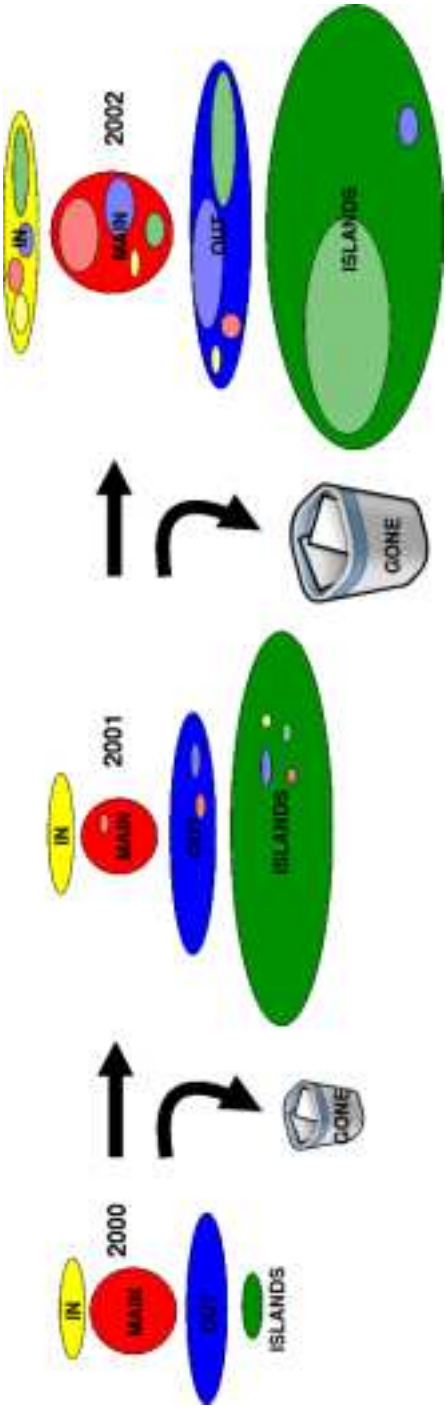

Figura 6: Flujo de los sitios a través de las componentes.

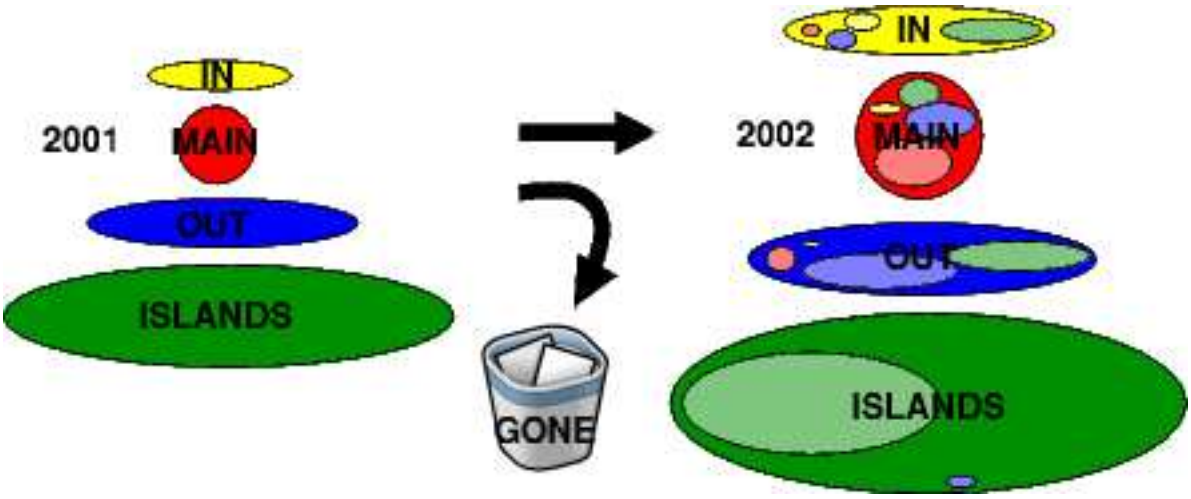

Figura 7: Flujo de los dominios a través de las componentes.

6. Las Consultas

En el presente capítulo se describe el análisis a las consultas realizadas por los usuarios del buscador de TodoCL. Se muestran ciertas distribuciones de éstas.

6.1. Frecuencia de palabras consultadas

Se observaron las frecuencias de consulta de las palabras en el buscador TodoCL. Las palabras más consultadas en el buscador TodoCL, en los períodos de agosto y septiembre del 2001, son las de la tabla 20 En esta tabla se descartaron artículos, preposiciones y otras palabras funcionales.

6.2. Frecuencia de consultas

A continuación haremos un análisis de la frecuencia con que son consultadas las palabras por los usuarios de TodoCL. En la siguiente figura se pueden apreciar las frecuencias con que son consultadas las palabras para los años 2001 y 2002. La figura 8 muestra la frecuencia de las palabras consultadas en el año 2002. Se observa que siguen una distribución tipo *Zipf* de parámetro 1.14 lo que es muy similar a lo observado elaño 2001, donde la frecuencia de las consultas seguía una *Zipf* de parámetro 1.4.

2001		2002	
Palabra	% consultas	Palabra	% consultas
CHILE	0.5 %	GRATIS	1.3 %
FOTOS	0.5 %	FOTOS	1.2 %
GRATIS	0.5 %	CHILE	0.9 %
SEXO	0.4 %	SEXO	0.7 %
HISTORIA	0.4 %	HISTORIA	0.6 %
MP3	0.3 %	ARGENTINA	0.6 %
VIDEOS	0.2 %	MP3	0.5 %
MUSICA	0.2 %	MEXICO	0.5 %
ARGENTINA	0.2 %	JUEGOS	0.5 %
LEY	0.2 %	MUSICA	0.4 %
UNIVERSIDAD	0.1 %	MANUAL	0.4 %
VENTA	0.1 %	DOWNLOAD	0.4 %
MEXICO	0.1 %	VIDEOS	0.4 %
SOFTWARE	0.1 %	SOFTWARE	0.4 %
INTERNET	0.1 %	LEY	0.3 %

Tabla 20: Palabras más consultadas en TodoCL.

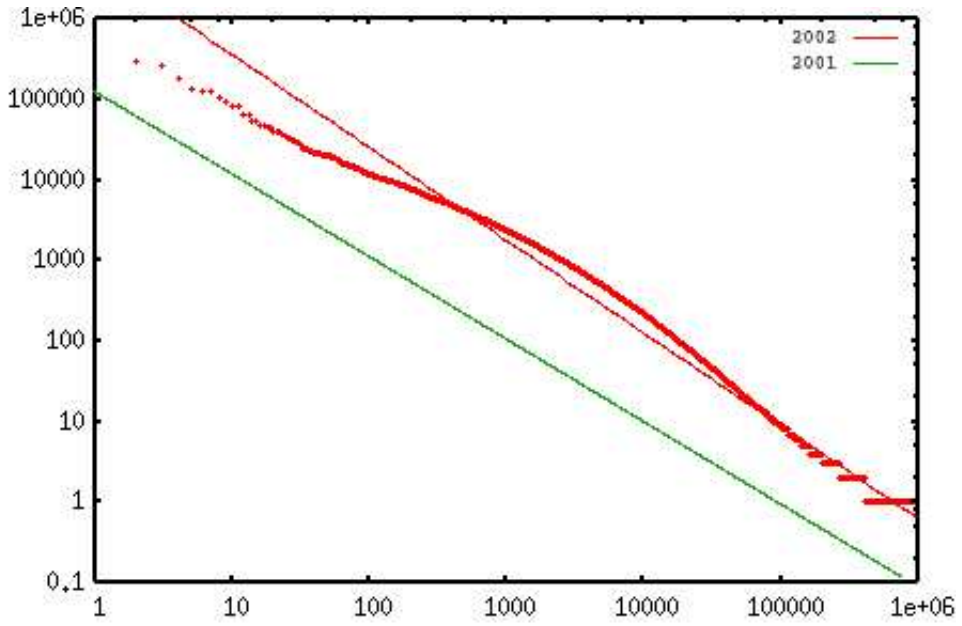

Figura 8: Frecuencia de consulta vs. Palabras.

6.3. Palabras consultadas y en el contenido

Las palabras consultadas y las que aparecen en las páginas siguen distribuciones similares. Surge la pregunta sobre su relación. En el gráfico de la figura 9 se ve la relación entre documentos relevantes y cantidad de consultas de las palabras. Lo más común son palabras con pocos documentos relevantes y pocas consultas. Hay palabras con pocos documentos y muchas consultas, ejemplos de esto son Hentai, México, DivX, Carátulas, y Melodías. Las palabras con muchos documentos relevantes y pocas consultas son, en general, preposiciones, pronombres y artículos como $pero$, $otros$, $este$, etc. Las palabras con mucho contenido y muchas consultas son, en general, $stopwords\ words$ como y, de, el y la; pero aparece de forma interesante $Chile$ como palabra muy consultada y que aparece en muchas páginas. Las palabras poco consultadas y con poco contenido no son interesantes, ya que son muchas. La relación de las palabras consultadas y las del contenido no es clara.

6.4. Opciones de consulta

Al utilizar un buscador, es posible alterar los parámetros bajo los cuales se realizará la consulta. Los parámetros existentes en los buscadores estudiados, en el modo de búsqueda simple, son:

Operador con valores AND, OR, FRASE. El valor AND busca documentos que tengan todas las palabras, OR documentos con alguna palabra, FRASE documentos que contengan la frase exacta.

Acentos considerar o no acentos en la consulta.

En la tabla 21 se pueden ver los niveles de utilización de cada opción en TodoCL. Los valores más altos, en ambos casos, son los valores por defecto. Esto le da una tremenda importancia a las opciones por defecto, ya que su elección será determinante, en una gran cantidad de casos, para el buen resultado de las consultas.

7. Conclusiones

A partir de este estudio es posible concluir diversos aspectos interesantes de la evolución en el tiempo de la Web Chilena. Desde el punto de vista de las cifras

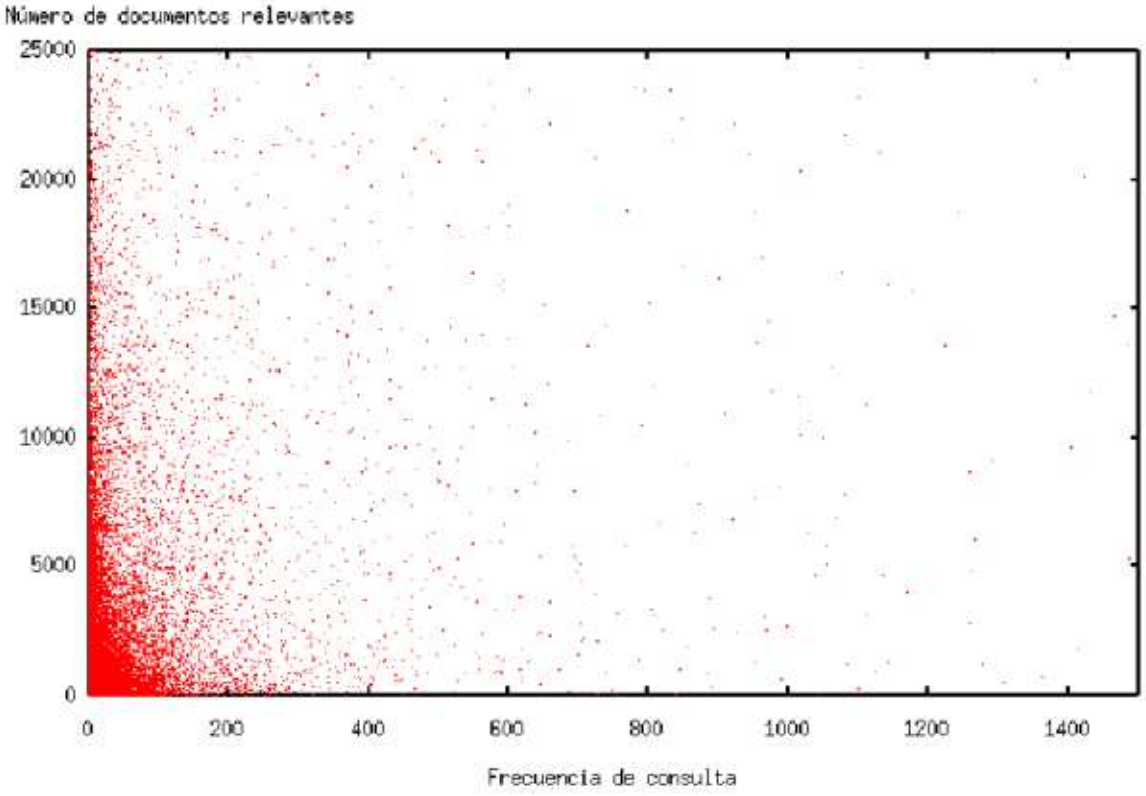

Figura 9: Cantidad de consultas v/s documentos relevantes para las palabras (2001).

Opción	% uso 2001	% de uso 2002
AND	99.9 %	84.5 %
OR	0 %	0.5 %
FRASE	0 %	15 %
con acentos	0.1 %	0.1 %
sin acentos	99.9 %	99.9 %

Tabla 21: Uso de las opciones de búsqueda en TodoCL.

globales lo más destacable es que el número de páginas, sitios y dominios presentes se ha prácticamente duplicado año a año. Esto refleja el crecimiento acelerado de la Web en Chile. Sin embargo, es importante señalar que el 56 % de los dominios y el 54 % de los sitios tienen sólo una página. Manteniéndose relativamente constante la distribución que siguen las páginas en la Web. También se observa que en general son siempre los mismos dominios los que poseen más páginas, no así los sitios.

Al analizar los medios y formatos de la Web en la actualidad llama la atención la importancia que ha tomado PHP dentro de las páginas dinámicas desplazando a ASP del primer lugar de preferencias.

En el periodo 2001-2002 se observa de forma especial las características de la componente ISLANDS, ya que los sitios que pertenecen a esta componente son los mayores en número. Los sitios en ISLANDS tienden a ser los más nuevos, lo que deja en claro el gran crecimiento de la Web. Esta componente es la que más ha crecido y a la vez la que más sitios ha perdido. También es interesante notar que MAIN y OUT son componentes altamente estables, manteniendo constante un porcentaje importante de sus sitios.

En cuanto a las palabras más consultadas se puede apreciar que estas no han cambiado mucho entre el 2001 y 2002, incluso la distribución que ellas siguen es prácticamente la misma.

Referencias

[1] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata,R., Tomkins, A. y Wiener, J., Graph structure in the Web. Proc. 9th International World Wide Web Conference (WWW9)/Computer Networks, 33(1-6),2000, pp. 309-320. Disponible en http://www9.org/w9cdrom/contents.html#CHARACTERIZATION.R.

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