Evolución de la Web Chilena 2001-2002 *

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

Depto. de Ciencias de la Computación
Universidad de Chile
Email: {rbaeza,bpoblete,fsaint}@dcc.uchile.cl

Resumen

Este es un estudio sobre las características de la Web Chilena, basado en los datos obtenidos por el buscador de páginas web TodoCL, en los años 2000, 2001 y 2002. Esta investigación recopila 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. Parte de los datos citados en este documento corresponden al primer estudio de la Web Chilena realizado en el año 2000. Esta información, combinada con datos obtenidos en forma reciente, permiten por primera vez realizar un análisis comparativo de la evolución de la Web Chilena en el periodo comprendido entre los años 2000 y 2002 a nivel de páginas, sitios y dominios.

1. Introducción

La Web se caracteriza por estar definida por un conjunto inusual y heterogéneo de elementos. Las mismas características que hacen de la Web un importantísimo medio de difusión y comunicación, hacen que su análisis sea muy complejo. Debido a esto, es que surge el interés y la importancia de realizar periódicamente una descripción de esta, tomando en cuenta sus propiedades más relevantes, su evolución y crecimiento en el tiempo. Dado el enorme tamaño de la Web en la actualidad, se hace fundamental la investigación de subconjuntos de esta. En el caso de este estudio se analizará el subconjunto correspondiente a la Web Chilena, a través de los datos recopilados por el buscador chileno TodoCL[1], parte del *spin-off* Barcino Ltda. del Departamento de Ciencias de la Computación de la Universidad de Chile, en colaboración con Akwan (Brasil).

Esta investigación recopila los datos recogidos por TodoCL en el periodo correspondiente a los años 2001 y 2002, lo cual permite llevar a cabo diversas comparaciones entre estos dos años y a su vez comparar estos resultados con los obtenidos en el primer estudio [2].

El análisis realizado se divide en dos 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. 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. Para finalizar se mostrarán conclusiones obtenidas del análisis de los datos expuestos en este documento.

*Parcialmente Financiado por Proyecto Fondecyt 102803 (CONICYT).

[1] Localizado en http://www.todocl.cl y donde se encuentra la versión completa de esta investigación

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.

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]. Como puntos de partida, para recorrer la Web, TodoCL utiliza principalmente páginas bajo el dominio .CL, más algunas páginas en el dominio .NET y .COM pertenecientes a empresas Chilenas.

Algunos conceptos importantes en cuanto a los buscadores son:

- **Colección**: Son todos los documentos recolectados e indexados por el buscador.
- **Página**: Es un documento indexado por el buscador.
- **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, de gran importancia para el análisis que se realizará en este documento, 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 que no pertenecen al dominio .CL, que se encuentran en Chile, más 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

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

Páginas	1988706
Sitios	38307
Dominios	34867

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

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.

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 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.

La mayor parte de las páginas que recolecta TodoCL pertenecen al dominio .CL. Las cifras exactas de páginas, sitios y dominios .CL encontrados el año 2002 por el buscador se presentan en la tabla 2.

En la tabla 3 se pueden apreciar datos sobre las ubicaciones físicas de los servidores en que se encuentran los sitios chilenos. El análisis de esta tabla se hizo según en número IP del servidor y su pertenencia o no a redes chilenas. Los números no son del todo consistentes con las cifras mostradas en las tablas anteriores, dado que para muchos sitios el DNS no respondió[5]. Es interesante ver la cantidad importante 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 la distribución de páginas puede ser aproximada por una *Zipf* con parámetro *1.58*. Esta distribución es muy similar a la observada para el el año 2001 (que también se puede apreciar en la figura 1), representada por una *Zipf* con parámetro *1.8*.

En comparación con el 2001, en que un 45 % de los sitios tenía sólo una página, 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 4 se presentan los sitios con mayor número de páginas en Chile para los años 2001 y 2002.

Al comparar los sitios presentados en la tabla 4 es posible observar que la totalidad de 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 aún estan entre los 1000 sitios con más páginas de Chile en el año 2002.

En la tabla 5 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 el caso de los sitios con más páginas. Es decir, la mayoría de los dominios que en el 2001 tenían el mayor número de sitios están presentes nuevamente en el 2002.

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

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

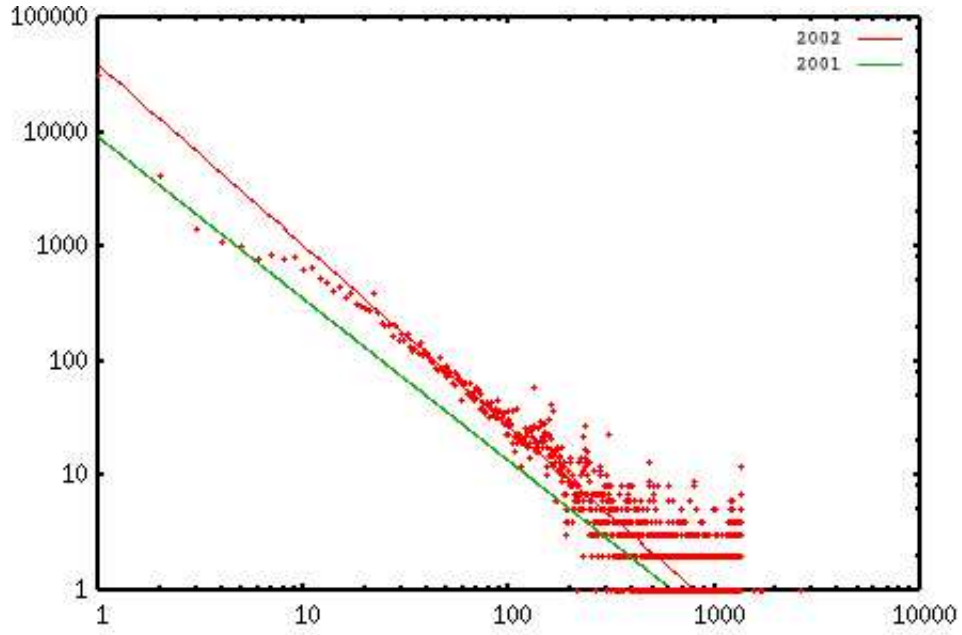

Figura 1: Cantidad de sitios vs. cantidad de páginas.

3.3. Tamaño

El tamaño de los sitios refleja el nivel de contenido que hay en ellos. 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. 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 6 se pueden ver los sitios con mayor contenido en tamaño en megabytes para el año 2001, considerando el tamaño del sitio completo, es decir, incluyendo archivos no indexables. En esta tabla también se puede apreciar el tamaño de los sitios medido en cuanto al texto plano, sin incluir archivos binarios ni tags HTML, de los sitios en el año 2002. De esta tabla cabe destacar que la mayoría de estos sitios corresponden a copias locales o *mirrors* del portal de software Tucows[6].

4. La Topología

En esta sección se describen elementos de la Web Chilena basados en las características topológicas de ésta, las cuales son principalmente, las páginas, sitios, 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 de ésta. Tanto así que la ubicación topológica de las páginas se ha considerado en Google como la primordial característica de jerarquización de los documentos, reemplazando a las más clásicas, basadas en distancia vectorial entre documento y consulta.

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

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.planoinmobiliario.cl	1355
www.codelco.cl	1268	foros.ircangol.cl	1355
www.australtemuco.cl	1266	www.redhat.cl	1354
www.sectormatematica.cl	1265	beta.ce.cl	1354

Tabla 4: Sitios con mayor número de páginas en Chile.

4.1. Enlaces

Los dominios más populares entre los distintos sitios web chilenos son los que se muestran en la tabla 7, la cantidad de referencias mencionadas es a nivel de dominios.

La tabla 8 muestra los sitios más referenciados a nivel de sitios.

5. Macroestructura

Es posible analizar la Web como un grafo, esto permite aplicar toda la teoría que existe sobre ellos 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 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 a los que se puede llegar desde 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 ninguna componente.

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 5: Dominios con mayor número de sitios en Chile.

Sitio(2001)	Tamaño(megabytes)	Sitio(2002)	Tamaño(megabytes)
pda.attla.cl	71588	www.plusvalia.cl	33783
www.embnet.cl	57821	www.anfitrion.cl	14522
www.linuxberg.cl	57003	www.oim.web.cl	14233
tucows.ctcinternet.cl	56137	www.camara.cl	14023
linuxberg.attla.cl	55857	c6.li2.uchile.cl	13247
tucows.rdc.cl	55228	sociales.uchile.cl	12848
tucows.firstcom.cl	54911	www.diariooficial.cl	12726
tucows.uplink.cl	48871	rehue.csociales.uchile.cl	12512
tucows.telsur.cl	47729	ads2.astro.puc.cl	11658
tucows.attla.cl	47149	www.csociales.uchile.cl	11295

Tabla 6: Sitios con mayor contenido en megabytes, incluyendo binarios (2001) y sólo texto (2002).

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 7: Dominios más referenciados a nivel de dominios.

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 8: Sitios más referenciados a nivel de sitios.

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

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 los sitios, aunque también 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 dedicados a diversos temas.

Al observar la tabla 9 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. Esto podría explicarse en la gran necesidad que se ha creado por contar con un sitio de presencia en Internet. Cada nueva institución u organización busca participación en este medio.

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

Tabla 10: Cifras generales de las componentes de la Web Chilena a nivel de sitios.

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

Tabla 11: Cifras generales de las componentes de la Web Chilena a nivel de dominios.

5.1. Composición de las Componentes Actuales

Desde los comienzos de la Web los sitios han cambiado su ubicación dentro de la topología de ésta. La pregunta que surge entonces es dónde están hoy los sitios de las componentes de hace un año. Como dato relevante se presentan en las tablas 10 y 11 algunas cifras globales de las componentes de la Web Chilena, a nivel de sitios y de dominios respectivamente. Estas tablas muestran la cantidad de sitios presentes cada año y la cantidad de sitios que desaparecen (GONE) y surgen (NEW) año a año.

En la tablas 12 y 13 se analiza el movimiento en las componentes a nivel de sitios y en la tabla 14 se hace el mismo análisis a nivel de dominios. Hay dos lecturas posibles de las tablas 12 y 13. 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 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 2 y 3 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 9, 10, 11, 13 y 14.

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

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 13: 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 14: Composición de las componentes a nivel de dominios en el 2002 respecto del 2001.

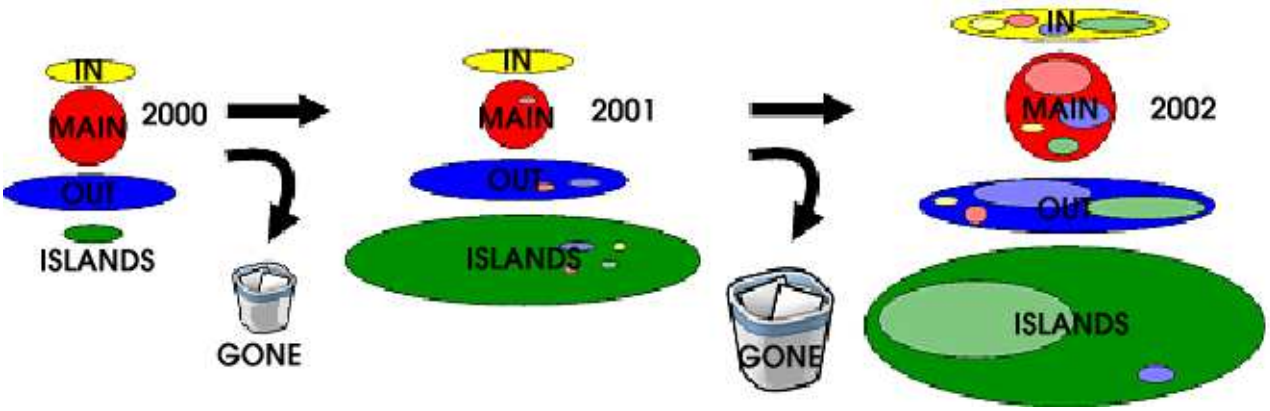

Figura 2: Flujo de los sitios a través de las componentes.

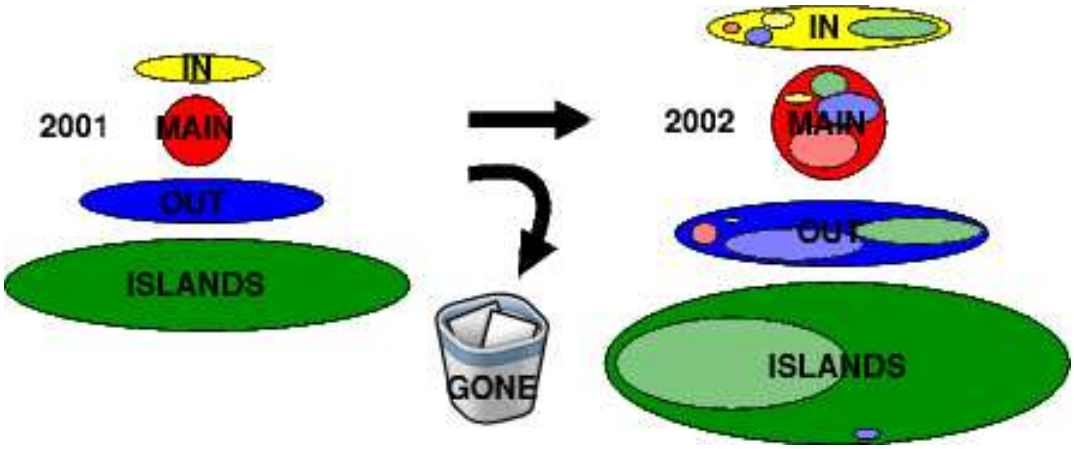

Figura 3: Flujo de los dominios a través de las componentes.

6. 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 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.

En el periodo 2001-2002 se destacan en forma particular 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. Esto habla de la importancia de los sitios en estas dos componentes, que al ser referenciados por muchos otros sitios implica que son importantes dentro de la estructura de la Web y que es muy probable que no sean sitios volátiles en el tiempo.

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.