Faz
revista de diseño de interacción

Comparación entre la Web Chilena y la Web Española

Eduardo Graells*, Ricardo Baeza-Yates**

Resumen

En este artículo se realiza una comparación entre las características de las webs nacionales de Chile y España, en base a los estudios previos dedicados a cada una de ellas. Dicha comparación considera el análisis de vocabulario,documentos, sitios, dominios, componentes estructurales, software utilizado por los servidores y proveedores de servicio. Se muestra que a pesar de ser dos webs diferentes comparten una gran cantidad de propiedades y similitudes.

1. Introducción

Se puede decir que Chile y España, a pesar de ser dos países de habla castellana, son muy distintos entre sí. No sólo es grande la distancia entre ellos: sus culturas tienen poco en común, sus tradiciones y su sociedad tienen diferencias bastante marcadas. En una época donde la globalización cada vez se expande más y la red internet se observa como un todo que a simple vista no distingue fronteras entre países, ¿se puede hablar de webs nacionales?. En este artículo se responderá esta pregunta, se verá que sí se puede hablar de una web nacional, en particular de las webs de Chile y de España. Esto se hará a través del análisis de las características de las páginas, sitios y dominios de cada país. También se estudiará el

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

contexto de ambas, mediante el análisis de los idiomas presentes en cada una, de los proveedores que hospedan, y las relaciones que se forman, mediante enlaces, en los diferentes sitios.

Para comenzar, la primera comparación objetiva que se puede realizar entre dos países se relaciona con sus datos geográficos y poblacionales. En la Tabla 1 se muestra un cuadro resumen donde se puede observar que la población de España es cerca de 2,7 veces la de Chile, mientras que la superficie de Chile es 1,5 veces la de España. La densidad de habitantes por kilómetro cuadrado de España es 4,1 veces la de Chile. Así mismo, una primera pregunta que se puede realizar es la siguiente: ¿se repiten estas proporciones en la Web de estos dos países?

	Chile	España	
Población	16.598.074	45.116.894	habitantes
Superfice	756.950	504.645	km2
Densidad	21,31	89,40	hab/km2

Cuadro 1: *Datos de población y superficie de ambos países.*

El resto del artículo está organizado de la forma siguiente. Primero explicamos la metodología y los conceptos básicos que se utilizan. Luego presentamos los resultados para las páginas, sitios y dominios. A continuación presentamos resultados sobre proveedores y servidores, para terminar con las conclusiones más importantes de la comparación.

2. Preliminares

2.1. Metodología

Para poder responder la pregunta principal formulada en la introducción, se deben explicitar las herramientas utilizadas y el criterio bajo el cual se han medido las diferentes características de la web. Como punto de partida en un estudio web se debe obtener la mayor cantidad de direcciones de sitios de esa web posible, de modo de recorrerlos, bajar las páginas web que contienen, y buscar en ellas nuevas direcciones para seguir recolectando sucesivamente; es un proceso que puede no terminar jamás, ya que la web puede ser considerada infinita y está siempre en expansión [2], pero existen algunos parámetros que ayudan a determinar cuándo es tiempo de detenerse. Los programas que realizan estas colectas son llamados *crawlers* ; el crawler utilizado en estos estudios es WIRE[1] .

Ahora bien, si se posee una lista inicial de sitios web y se encuentran enlaces a otros sitios nuevos, ¿cómo determinar si una dirección pertenece a una web nacional? Se pueden adoptar muchos criterios, pero el más adecuado según los estudios realizados con anterioridad es considerar como web asociada a un país todos los sitios que se encuentran hospedados en direcciones IP dentro del rango asociado a ese país. Quedan fuera, por ejemplo, los sitios que se encuentran hospedados en el extranjero pero que son de personas o instituciones del país en estudio.

Las direcciones iniciales utilizadas en los estudios, llamadas *semillas*, son, en el caso de Chile, todos los dominios .cl que existen al momento de hacer el estudio, gracias a un acuerdo con NIC Chile, más sitios con otros dominios que hemos obtenido del buscador TodoCL[2]. En España se utilizó el directorio Buscopio[3] para obtener las primeras semillas; en este caso no basta con contar con la lista de dominios .es ya que eran poco utilizados en esa época por las condiciones que se pedían para inscribir un dominio, las cuales ya se han simplificado.

Una vez que ya se tienen las colectas con los sitios y páginas que queremos estudiar, ¿cómo se pueden comparar? No basta con comparar las características básicas de las colectas, como el número de sitios y el promedio de páginas por sitio, es necesario establecer una relación que entregue datos más significativos. En esto tiene gran importancia saber que la web global es una *red libre de escala* [6], lo que quiere decir que una

[1] Web Information Retrieval Environment, desarrollado en el Centro de Investigación de la Web, http://www. cwr.cl.
[2] http://www.todocl.cl
[3] http://www.buscopio.cl

muestra más pequeña de ella mantiene propiedades de la muestra completa. Bajo este esquema, las webs de Chile, Espãna y de otros países, a pesar de ser muy diferentes en su planteamiento y en su contenido, debieran ser similares desde un punto de vista analítico.

No todas las propiedades de la web se pueden comparar directamente mediante porcentajes o proporciones. Un punto de comparación usado es la aproximación de diferentes características a una ley de Zipf, llamada así en honor a George Kingsley Zipf. En 1932 modeló la distribución de la frecuencia de palabras en los textos, que resultó ser muy sesgada: algunas palabras son utilizadas con mucha frecuencia mientras que otras raramente lo son. Este mismo comportamiento se puede observar en las redes libres de escala, en particular con los enlaces entre distintos nodos: algunos nodos acaparan todos los enlaces mientras que otros reciben muy pocos. También, en 1896, VilfredoPareto observó este comportamiento cuando modeló la distribución de la riqueza: el 80% de la riqueza está repartida en el 20% de la población.

En términos matemáticos, una Ley de Zipf es una distribución de datos que sigue una ley de potencias, es decir, la probabilidad de encontrar un elemento de tamaño x es proporcional a una potencia, siendo ésta, en una ley de Zipf, χ^{-a} . El valor α es llamado *parámetro de la distribución*, y es el punto de comparación utilizado. Dos distribuciones pueden tener un valor para x diferente, pero si tienen un parámetro similar entonces observamos propiedades comunes. A modo de ejemplo, en la Figura 1 se muestra la distribución del vocabulario de la web Chilena. En el eje x las palabras han sido numeradas a partir de 1, ordenadas desde más frecuente a menos frecuente, en el eje y se observa la fracción o porcentaje en todo el vocabulario que corresponde a esa palabra.
La distribución del vocabulario se ha representado en escala logarítmica. Se observa que se puede trazar una línea recta que representa a gran parte del vocabulario: esa línea es la aproximación a una ley de Zipf. El parámetro corresponde a la pendiente de esa recta multiplicada por –1.

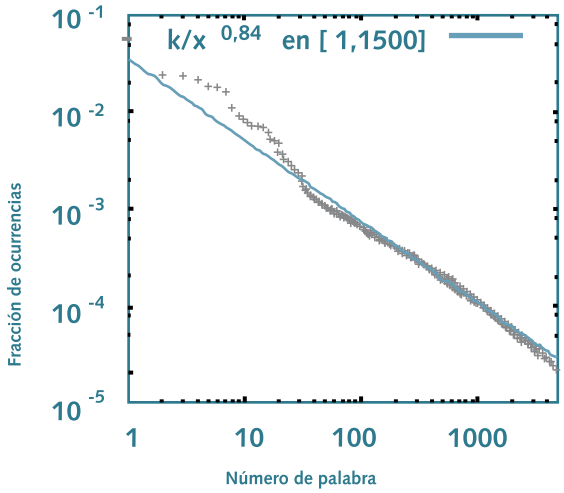


Figura 1: *Ejemplo de una distribución de Zipf: palabras en un texto.*

2.2. Datos Utilizados

La Tabla 2 muestra los datos generales de las webs de Chile [4] y de España [5]. Se aprecia que, del mismo modo que con los datos poblacionales y geográficos, en términos cuantitativos España supera a Chile en el número de páginas, sitios y dominios, lo cual tiene sentido si se considera que la población es mucho mayor. A pesar de ello, la proporción entre los datos (comparables a la densidad de habitantes) no es tan distinta: un sitio web en Chile promedia 43 páginas mientras que un sitio web en España promedia 52. Por otro lado, Chile tiene más dominios inscritos, pero muchos menos sitios web, lo que indica que en ambos países los dominios reciben usos muy diferentes.

Es necesario indicar que de estas cantidades presentadas no son totalmente exactas: algunas dependen de la configuración del crawler utilizado y del espacio disponible a la hora de hacer las distintas colectas[4] ; otras son sensibles a diferentes fenómenos que se pueden encontrar en la web, desde el SPAM o la generación de páginas dinámicas que no pueden ser detectadas.

[4] Por ejemplo, el límite de tamaño para las páginas recolectadas suele ser entre 100 y 110 kilobytes, lo que es suficiente para casi todas las páginas. Asímismo, de cada sitio se suele definir un límite entre 10 y 15 mil páginas para descargar como máximo, lo cual siempre ha sido suficiente.

Páginas web	Chile(2006) 7.403.840	España 2005 6.171.267	
Texto en Total	48,56	43	GiB
Texto promedio por página	7,04	2,78	KiB
Sitios Web	171.213	308.822	
Páginas promedio por sitio	43,24	52,08	
Texto promedio por sitio	304,59	146	KiB
Dominios	158.853	118.248	
Sitios promedio por dominio	1,08	2,61	
Páginas promedio por dominio	46,61	136,75	
Texto promedio por dominio	328,29	373	KiB

Cuadro 2: *Datos utilizados de la Web de Chile y España.*

3. Características de los Documentos

3.1. URLs y títulos

Al hablar de documentos se hace un análisis de cada página web sin agrupar por el sitio o dominio al que pertenece. El primer dato relevante es el largo de la URL mediante la cual se accede a cada uno de ellos, que en promedio para Chile es de 71 carácteres; para España, 67. Los promedios no sólo son similares sino que las distribuciones de los largos para ambos países se pueden aproximar por una distribución normal de parámetros similares. Una característica muy importante de las páginas, tanto para los recolectores como para los usuarios, son los títulos de ellas. Una de las primeras cosas que ve un usuario al abrir una página es su título en la barra de título del navegador; también es lo que más se destaca en los resultados de búsqueda de algún buscador, y el nombre con el que queda guardada en la lista de *Bookmarks o Favoritos*. Sin embargo, pese a su importancia, no todos los documentos tienen un título adecuado. En la Tabla 3 se observa la proporción entre los tipos de título encontrados en las webs estudiadas. Lo recomendable es tener un título único para cada documento, aunque es común que un sitio tenga varios documentos con un título compartido. Los títulos por omisión son aquellos del tipo *"Página nueva"* o *"Página sin título"*, característico de los programas de diseño WYSIWYG.

Es difícil para un sitio grande lograr tener un título único para cada documento, pero sin duda alguna el esfuerzo tiene sus recompensas: en particular, los usuarios podrán distinguir fácilmente entre resultados de un buscador, y muchas veces el resultado que eligen depende en gran parte del título que se les presenta. Un buen título debería representar documento e indicar el sitio al que pertenece.

	Chile	España	
Compartido	51,12	71,33	%
Único	33,22	15,85	%
Vacío	12,1	9	%
Por omisión	2,47	3	%

Cuadro 3: *Distribución de títulos en páginas.*

Al medir los largos de los títulos se estimó que el 95% de los títulos de España no supera los 25 carácteres. En el caso de Chile los títulos se e cuentran prácticamente repartidos de manera uniforme entre los 20 y 60 carácteres. Esto no quiere decir que todos los títulos de la web chilena sean más descriptivos, ya que se encontraron muchos títulos que, si bien tenían un largo considerable, sólo contenían carácteres con fines decorativos.

3.2. Texto de los documentos: tamaño, idioma y vocabulario

Por texto de los documentos se entiende tanto al contenido leíble, con significado para un humano, como al contenido no leíble, es decir, todo el contenido del documento que el usuario no lee pero que de alguna manera afecta al documento, como pueden ser etiquetas HTML para el formato, código javascript para interacción y estilos CSS incrustados en el documento para la presentación. No se consideran archivos adjuntos o enlazados, como imágenes, vídeos, audio, o incluso archivos con código CSS o javascript; de ellos sólo guardamos un registro. Considerando esto, y comparando el tamaño de los documentos y la fracción de ellos que tiene ese peso específico, se puede observar una distribución que sigue una ley de Zipf en su parte central. Los parámetros encontrados son 2,64 para Chile y 2,25 para España, lo cual indica que hay muchos documentos que pesan poco y pocos documentos que pesan mucho.

Respecto al texto leíble dentro de los documentos, en Chile el 80% de las páginas se encuentra en castellano, con prácticamente todo el resto de los documentos en inglés: otros idiomas, como el francés y el alemán, tienen una presencia mínima, inferior al 1%. Esto no es así en el caso de España, donde el castellano tiene una presencia del 52,38% del total de documentos; el inglés, de 30,27%; el catalán, de 8,18%; el francés, de 5,89%; y el Alemán, de 1%, repartiéndose el escaso porcentaje restante entre otros idiomas. En general, la lengua oficial suele ser la que tiene mayor presencia en una web nacional. Sin embargo, en algunos países esto no ha sido así, es el caso de Tailandia, donde el inglés tiene mayor presencia que el tailandés. Indudablemente el comercio y el turismo tienen mucha influencia en estas cifras [3].

Las diez palabras más comunes en la web, eliminado artículos, adverbios y otras palabras que no tienen significado[5] , son:

Chile: chile, producto, usuario, todos, servicio, mensaje, empresa, comentario, web, santiago.

España: artículo, información, trabajo, ley, servicio, madrid, año, universidad, forma, españa.

No deja de llamar la atención que entre las palabras más usadas se encuentre el nombre del país y su capital, y que algunas palabras se repiten, mientras que otras son sinónimas o tienen relación entre sí.

3.3. Tecnologías

Para los usuarios la tecnología que genera a una página web es indiferente, pero desde el punto de vista del recolector esto no es así, ya que no es lo mismo una página estática, existente en el servidor hasta que alguien la borre, que una página dinámica, que se genera al momento de acceder a ella. Una página dinámica no tiene una existencia mayor al instante en que fue accedida mediante una dirección que la identifica, aunque ciertamente si se accede a la misma dirección entregando los mismos parámetros probablemente se obtenga un documento con los mismos contenidos.

Es importante poder detectar si una página es dinámica, porque entre ellas se pueden establecer recursiones que lleven a un número infinito de páginas dentro de un sitio web. Al momento de llevar a cabo la recolección, un 42,5% de las páginas chilenas y un 22% de las páginas españolas fueron identificadas como páginas dinámicas. Al inspeccionar los sitios web se puede observar que estas cifras en la realidad son mucho mayores.

Existen varias tecnologías que permiten generar páginas dinámicas, siendo las dos más usadas a nivel mundial PHP y ASP. PHP es una tecnología de código abierto y de uso gratuito; ASP es tecnología cerrada y sólo se puede utilizar con servidores Microsoft IIS. En Chile el 75% de las páginas dinámicas es generado mediante PHP, un 21,4% es generado mediante ASP, y el resto por otras tecnologías menos usadas como JSP

[5] Este tipo de palabras se conoce como stopwords o palabras funcionales, y corresponden a las palabras que tienen la mayor frecuencia en la distribución.

o Cold Fusion. En España PHP también supera a ASP, pero no por mucho: 46,24% versus 41,65%.

3.4. Documentos que no están en HTML

Existen diversas aplicaciones y formatos para trabajar con documentos, de cierta forma se puede establecer un orden de popularidad de acuerdo a la presencia de sus formatos en la web. Se encontró una gran cantidad de documentos en formatos distintos a HTML[6]: en ambos países los documentos en formato PDF tienen una gran presencia, en particular en Chile donde son los documentos con mayor participación, con un 53% del total de documentos. En España la participación de PDF es de un 41,26%, siendo superado levemente por los archivos de texto plano, TXT, con un 41,68%. A su vez, TXT no tiene ese mismo nivel de presencia en Chile, pues sólo un 13% de los documentos estaban en ese formato. El segundo formato más frecuente en Chile, con un 22% de participación, es XML, es decir, documentos de diferentes propósitos pero que guardan su información utilizando XML. El resto de los documentos, que en ambos países rodea al 17%, se reparte entre archivos de la suite Microsoft Office, con una mínima participación de otros formatos como PostScript u OpenOffice.

3.5. Enlaces entre páginas web

El número de enlaces que recibe un documento puede tener relación con su importancia o su popularidad en la web, y como era de esperarse, se encontró un gran desbalance entre los que tienen muchos enlaces y los que tienen pocos.

El grado interno de un documento, es decir, el número de enlaces que recibe, se puede distribuir con una ley de Zipf con parámetro 1,95 en Chile y con parámetro 2,11 en España. Es interesante saber que en Chile el 75% de los documentos posee todo el grado interno, es decir, un 25% de los documentos no tiene ningún enlace hacia él en la web nacional que le corresponde, lo que no quiere decir que no reciba enlaces desde la web global.

El grado externo, o el número de enlaces que posee un documento hacia otros, también puede ser representando mediante una ley de Zipf, de parámetros 3,51 para Chile y 2,84 para España. Indudablemente estas distribuciones indican una desproporción muy grande entre los documentos con pocos enlaces externos y los que tienen muchos enlaces externos. En general los documentos con muchos enlaces no son generados por humanos, sino más bien corresponden a directorios generados automáticamente. En Chile el 45% de los documentos posee todo el grado externo.

A cada documento se le puede calcular un valor que represente su popularidad en términos de enlaces en la web. El valor más conocido, tanto en la comunidad científica como en los usuarios, en particular los que tienen sitios, es PageRank [7]. PageRank considera el número de enlaces que llega a un documento, aunque también toma en cuenta la posibilidad de que una persona llegue de forma aleatoria a él (por ejemplo, a través de un enlace en sus favoritos), por lo que incluso un documento que no ha sido enlazado nunca tiene un PageRank distinto de cero. Habiendo calculado PageRank para las páginas de ambos países, dentro del contexto de la web nacional, se observa que en general su distribución, es decir, un valor de PageRank y la fracción de los documentos asociados a él, se puede expresar mediante una ley de Zipf de parámetro 2,09 en Chile y 1,96 en España. El valor de este parámetro para la web global es de 2,1.

4. Características de los Sitios

La noción de sitio web que maneja un crawler es diferente a la que maneja un usuario. Por ejemplo, dentro del departamento de una universidad usualmente cada profesor tiene su propio conjunto de páginas web in-

[6] Estos documentos no son descargados. Se guarda un registro de ellos.

dicando sus actividades académicas, publicaciones y proyectos. A ojos de un usuario el conjunto de páginas de un profesor es un sitio web, y el conjunto de páginas de otro profesor es un sitio diferente. A ojos de un crawler ambos sólo son un conjunto de páginas pertenecientes al sitio de la universidad.

En base a esto, un sitio web se define como el conjunto de documentos que comparten la parte de la URL que identifica al servidor que lo contiene. De este modo www.dimec.uchile.cl y www.dcc.uchile.cl son dos sitios diferentes. Además se considera la heurística que dice que www.sitio.es y sitio.es (sin el *www*) son el mismo sitio.

4.1. Documentos en los sitios

La distribución de documentos en los sitios también se puede modelar mediante una ley de Zipf, de parámetro 1,74 para Chile y de parámetro 1,14 para España. Los sitios con mayor cantidad de páginas se pueden apreciar en la Tabla 4.

Chile		España	
Páginas	**Sitio**	**Páginas**	**Sitio**
13654	graphologychile.cl	12918	europages.es
11571	upadiseno.cl	12756	iei.ua.es
10083	joomla.gsuez.cl	12043	andalucia junta.es
9607	tabanotv.cl	11855	virtual.usc.es
9471	cepal.cl	11819	dei.inf.uc3m.es
9032	eclac.cl	11603	cvc.cervantes.es
8900	vmf.cl	11016	mundial2002.terra.es
8752	conciencia-animal.cl	10838	ftp.gui.uva.es
8538	directorioweb.cl	10560	edu.aytolacoruna.es

Cuadro 4: *Sitios con más páginas.*

Muchos de ellos, y de otros que también tienen una gran cantidad de páginas, no son tan grandes como parecen. Esto se debe a que existen diferentes motivos por los cuales una página se descarga varias veces sin detectar que ya ha sido descargada, porque se presenta como una página distinta a la ya conocida. Entre estos motivos se encuentran los sitios con URLs mal formadas, los que generan páginas dinámicas para cualquier URL que se entregue, los que generan páginas dinámicas y le entregan los parámetros (por ejemplo, el nombre de una sección o una consulta) en la dirección misma y no después del signo ? como dice el estándar. Esto se puede entender como sigue:

Correcto: http://es.search.yahoo.com/search?p=estudio+web

Incorrecto: http://es.search.yahoo.com/search/estudio/web

Se puede observar que la web Española, a pesar de tener muchos más sitios y documentos, los más grandes en cantidad de documentos tienen una magnitud similar a los sitios chilenos. Sin embargo, la mayoría de los sitios chilenos en la tabla sufren alguna anomalía como las indicadas anteriormente, mientras que varios de los sitios españoles de la tabla efectivamente disponen la cantidad mencionada de páginas.

4.2. Tamaño

Si se agrupan los documentos por sitio y se vuelve a estimar una representación del tamaño, se obtiene que en Chile el tamaño en *megabytes* de un sitio y la frac-

ción de sitios que posee ese tamaño se puede representar mediante una ley de Zipf de parámetro 1,57; en España se puede representar con una ley de Zipf de parámetro 1,15. En la Tabla 5 se muestran los sitios con mayor cantidad de texto.

Chile		España	
Texto[MiB]	Sitio	Texto[MiB]	Sitio
418	almacenesparis.cl	165	cortesclm.es
401	lanaciondomingo.cl	142	maia.ub.es
394	lnd.cl	140	blues.eurovia.es
388	almacenes-paris.cl	126	constitucion.rediris.es
386	diariolanacion.cl	123	rfc.imasd.elmundo.es
378	fo.cl	113	srftp.usc.es
370	bookings.cl	112	ftp.usc.es
369	booking.cl	109	senado.es
354	lanacion.cl	100	genome.imim.es
348	concilio.cl	99	cidob.es

Cuadro 5: *Sitios con más contenido.*

La magnitud en tamaño de los sitios no parece ser similar. Dentro de los sitios más descargados de Chile se observa por inspección que el gran tamaño se puede deber a las anomalías indicadas anteriormente. Además varios de esos sitios están repetidos, con diferentes dominios, pero también con diferentes tamaños. En la web chilena los sitios con mayor tamaño en su mayoría son catálogos de productos y algunos medios de comunicación. En la web española destacan sitios del gobierno, documentación de software y sitios universitarios.

4.3. Enlaces entre sitios

Los enlaces entre los sitios también se distribuyen de acuerdo a una ley de Zipf. El grado interno de los sitios se puede estimar con un parámetro de 1,99; para España el valor de este parámetro es 1,82. El grado externo se estima con parámetros de 1,91 y 1,34, respectivamente. Los sitios más enlazados se muestran el la Tabla 6:

Entre los sitios más enlazados de Chile se encuentran sitios del gobierno y sitios universitarios en los primeros lugares. En España los primeros lugares son de un sitio comercial, de un medio de comunicación, y del gobierno, respectivamente. Los sitios con más enlaces hacia otros sitios se pueden ver en la Tabla 7.

En este listado de sitios se encuentran más sitios comerciales, aunque también se pueden encontrar directorios y buscadores como TodoCL e Hispavista.
Respecto al PageRank, se puede sumar el PR de todos los documentos dentro de un sitio con el fin de obtener la suma de PageRank total. En este caso nuevamente se obtiene una ley de distribución de Zipf para la distribución del PageRank en los sitios, con parámetros 1,05 para Chile y 1,76 para España

4.4. Macrocomponentes y Estructura de la Web

Al igual que la web global, cada web nacional es un grafo dirigido. Dentro de un grafo se dice que una de sus partes es fuertemente conexa si es posible ir desde cualquier nodo (en este caso, un sitio) a otro dentro de esa misma parte. Se dice que esa parte es fuertemente conexa si se respeta la dirección de los enlaces, esto quiere decir que dentro de una componente fuertemente conexa es posible ir desde un sitio hasta otro sólo siguiendo enlaces entre sitios. Por supuesto que no toda la web es fuertemente conexa, ya que algunos sitios no contienen enlaces salientes, otros no tienen

Chile		España	
Enlaces entrantes	Sitio	Enlaces entrantes	Sitio
1192	sii.cl	1312	adobe.es
963	uchile.cl	1153	elpais.es
877	mineduc.cl	1128	boe.es
804	meteochile.cl	992	terra.es
680	bcentral.cl	977	rediris.es
659	puc.cl	974	mec.es
624	corfo.cl	969	ucm.es
600	sernatur.cl	956	csic.es
598	latercera.cl	883	abc.es
589	terra.cl	863	mcyt.es

Cuadro 6: *Sitios con más enlaces a ellos.*

Chile		España	
Enlaces hacia otros sitios	Sitio	Enlaces hacia otros sitios	Sitio
4480	todocl.cl	6418	aui.es
3337	3tetra.cl	5164	guia.hispavista.es
2075	compraseguro.cl	3445	sol.es
1772	bingos.cl	1647	personal4.iddeo.es
1718	boom.cl	1636	congreso.es
1449	portalciudadano.cl	1615	inicia.es
1197	yes.cl	1473	universia.es
908	huellas.cl	1440	terra.es
870	fotolog.cl	1408	grn.es
761	buscamos.cl	1270	personales.mundivia.es

Cuadro 7: *Sitios que contienen más enlaces.*

enlaces entrantes, e incluso hay sitios totalmente aislados.

Existen muchas componentes fuertemente conexas, desde las más sencillas que tienen un único componente, hasta una componente gigante, que contiene muchos sitios más que las otras componentes. La presencia de esta última es típica en redes libres de escala. Si consideramos solamente los sitios que tienen al menos un enlace saliente o un enlace entrante, se observa que en Chile el 14% de los sitios se encuentran en la componente gigante. En España el porcentaje de sitios en la componente gigante es del 15%, es decir, en ambos países una proporción similar de sitios está fuertemente conectado entre sí.

La componente fuertemente conexa gigante puede ser utilizada como un punto de partida para distinguir ciertas componentes estructurales de la web. Estas componentes se han definido como:

MAIN, sitios que están contenido en la componente gigante.

OUT, sitios que son alcanzables desde MAIN, pero que no tienen enlaces hacia ella.

IN, sitios que pueden alcanzar a MAIN, pero que no tienen un enlace hacia ellos desde MAIN.

ISLAS, sitios que no son accesibles ni hacia ni desde MAIN.

TENTÁCULOS, sitios que sólo se conectan con IN u OUT, pero en el sentido inverso, es decir, reciben enlaces desde IN u OUT.

TÚNEL, una componente que une OUT con IN sin pasar por MAIN.
Dentro de MAIN también existen otras sub-componentes, pero su comparación se ve con más detalle en los estudios correspondientes. En Figura 2 se puede observar la proporción de los sitios que pertenecen a cada componente definida previamente en las webs de Chile y España.

Gran parte de los sitios pertenecen a la componente ISLAS (49,49% en Chile y 81,63% en España). A pesar de ello, la mayoría de los documentos se encuentra en la componente MAIN (53% en Chile y 41,31% en España).

4.5. Sitios de 1 página

Al hablar de los sitios con mayor cantidad de documentos, se mencionaron algunas anomalías que distorsionan las páginas recolectadas de un sitio. Ahora bien, también se puede dar la situación en la cual un sitio no puede ser recolectado por completo. Esto puede ser natural en el caso de un sitio privado, donde se requiere un nombre de usuario y una contraseña; bajo esa condición sólo se puede recolectar una página, la que tiene el formulario de ingreso. También se da el caso de sitios que efectivamente tienen una única página: sitios en construcción, sitios *placeholders*, que sólo reservan el dominio, o sitios que sólo sirven para redireccionar al usuario a otro sitio, entre otros. Pero existen sitios que tienen páginas públicas que deberían poder recolectarse, y sin embargo no se puede bajar más que la primera de sus páginas, probablemente la portada. Estos sitios tienen solamente una página visible para los recolectores porque utilizan tecnología que depende del usuario para poder visitar las páginas restantes; ejemplos de esto son sitios que utilizan javascript o Flash para su navegación. Esto no sólo incide en estos estudios, también en la popularidad de esos sitios porque los recolectores de los buscadores no pueden indizarlos.

En la Web de España se encontraron 184.015 sitios de una página, es decir, el 60% de ellos. En Chile, los

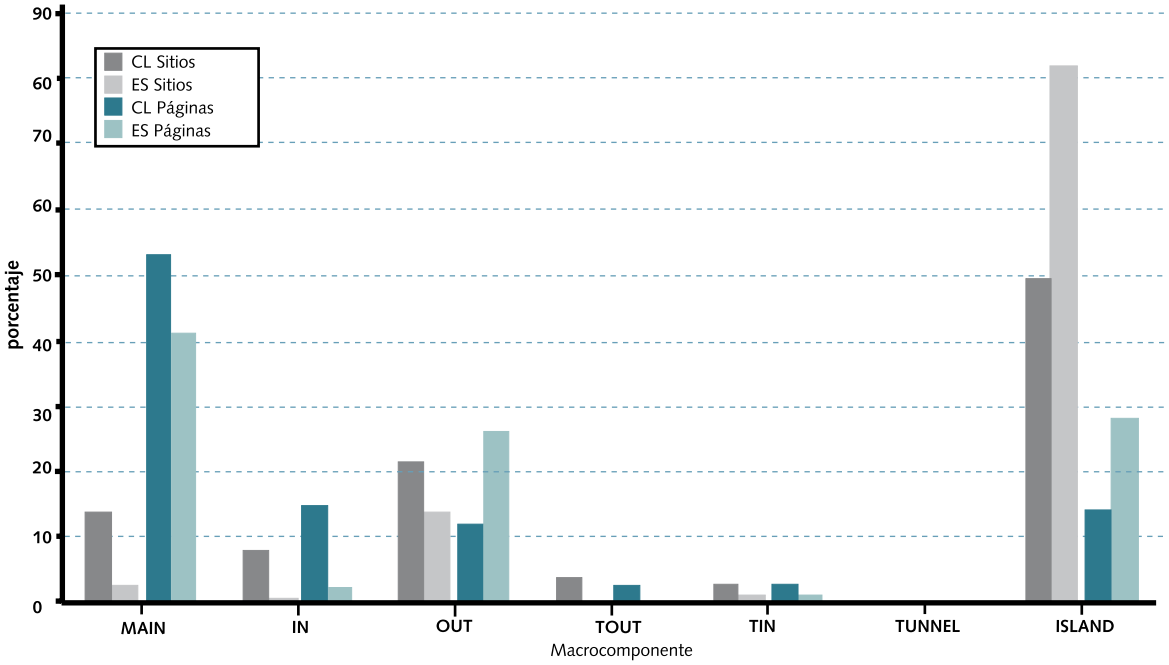


Figura 3: *Comparación entre software utilizado en los servidores.*

sitios de una página encontrados fueron 36.654, cerca del 21% de los sitios. Es necesario destacar que de los sitios de una sola página, en ambas webs, cerca del 50% son islas, es decir, sitios aislados de los demás. El resto pueden ser sitios normales cuya recolección no se realizó por los motivos ya indicados, o incluso sitios que realmente contienen una única página pero que de todos modos tienen contenido y han sido enlasados por otros.

5. Características de los Dominios

5.1. Número de Sitios por Dominio

En Chile existen 155.784 dominios con un sitio, aunque varios dominios superan con creces el promedio de 1,08 sitios por dominio. En España son 111.415 los que tienen solamente un sitio, habiendo 30 dominios con más de 1.000 sitios cada uno, superando el promedio de 2,55 sitios por dominio. En ambos casos la distribución de sitios por dominio sigue una ley de Zipf, de parámetros 1,22 y 1,23 respectivamente. En la Tabla 8 se muestran los dominios con más sitios.
En Chile se observan dominios comerciales, de gobierno y universitarios. En España, sin embargo, se observan muchos dominios que tienen solamente sitios de una página (son aquellos en los cuales el número de sitios es igual al número de páginas), o al menos una proporción casi 1 a 1 entre sitios y páginas. Estos dominios emplean una técnica llamada *DNS comodín* (*DNS wildcarding* en inglés), en la cual entregan una dirección IP válida y una página con contenidos sin importar la dirección solicitada. Se puede solicitar una dirección x.dominio.es, siendo x realmente inexistente en el servidor, pero recibir una página válida de todos maneras. Estas páginas probablemente sólo contienen publicidad y enlaces a sitios con SPAM.

5.2. Número de páginas por dominio

El promedio de páginas por dominio es de 47 para Chile y 133 para España. La distribución de páginas por dominio y fracción de ellos se puede representar mediante una ley de Zipf de parámetros 1,67 y 1,18, respectivamente. Además, en Chile hay 34.810 dominios con una sola página, lo que representa cerca del 21% de los dominios, muy similar a la proporción de sitios con una sola página. En España esto no es así, ya que hay 32.008 dominios con una sola página, representando al 26% de los dominios. Esto es lógico si se piensa que comprar un dominio tiene un costo asociado, mientras que una vez que se compra el do-

Chile			España		
Sitios	**Páginas**	**Dominios(.cl)**	**Sitios**	**Páginas**	**Sitio**
1407	3641	portalciudadano	24886	25006	bcnlink.com
499	86174	uchile	19118	19810	totanuncis.com
459	17052	boonic	5785	5785	onlinegamess.com
340	2571	scd	4865	4865	downloaddownload.co.uk
329	74542	terra	4641	4641	programasprogramas.com
308	10083	ucn	4637	4637	partituraspartituras.com
285	117	notarial	3951	3951	spiele-pc.com
249	3331	co	3817	3817	recetascocinar.com
156	23083	gov	3323	3323	juegosplus.com
146	25019	utfsm	3304	3304	masjuegosonline.com

Cuadro 8 *Dominios con más sitiose*

minio, agregarle nuevos sitios tiene costo cero.

5.3. Tamaño total de los dominios

En promedio, un dominio en Chile tiene un tamaño total de 328 KiB; en España el promedio es superior: 373 KiB. En ambas webs se puede aproximar la distribución de los dominios por una ley de Zipf de parámetros 1,28 y 1,19, respectivamente. En la Tabla 9 se muestran los dominios más grandes en tamaño.

Chile			España		
Tamaño[MiB]	Dominio	Tipo	Tamaño[MiB]	Dominio	Tipo
2131	decompras	C	1939	europages.es	C(S)
1782	buy7	C	1939	upm.es	U
1313	qsale	C	532	upc.es	U
1173	uchile	E	518	rediris.es	I
1107	terra	C	491	csic.es	I
900	k21	C	473	iespana.es	C(S)
648	deremate	C	448	elmundo.es	C
603	mercadolibre	C	427	ua.es	U
569	canal13	C	401	uvigo.es	U
480	laguiachile	C	381	usc.es	U

Cuadro 9: *Dominios de mayor tamaño*
C: Comercial, E: Educativo, I: Investigación, U: Universidad, C: Comercial, S: SPAM.

En Chile los dominios más grandes son los de remates o de catálogos, ya que tienen un gran número de páginas con información redundante sobre los productos, y muchos de ellos se copian el contenido entre sí. En España los sitios más grandes son en su mayoría universitarios; salvo por un sitio, en la lista presentada el SPAM está ausente. Esto se puede deber a que el SPAM contamina la red con enlaces y páginas falsas o publicitarias, pero no con el suficiente contenido para superar en tamaño a un sitio normal.

5.4. Dominios más enlazados

En la Tabla 10 se muestran los dominios más enlazados para cada país. En el caso del dominio chileno que recibe más enlaces es muy probable la presencia de SPAM o de alguna anomalía. El dominio español que recibe más enlaces es adobe, debido a la gran cantidad de enlaces para descargar el software *Adobe Reader*.

5.5. Dominios de Primer Nivel

En la Tabla 11 se muestra la distribución de sitios y páginas en los diferentes dominios de primer nivel al que pertenecen los sitios en estudio. La primera fila corresponde al dominio nacional, es decir, cl para Chile y es para España. Se omitieron otros dominios cuya presencia correspondía a un número demasiado pequeño de sitios o páginas.

Se observa que en Chile el dominio más utilizado es el nacional, mientras que en España sólo un 15,9 % de los sitios usa el dominio es[7] , aunque ese porcentaje de sitios alberga el 56 % de las páginas. Los domi-

[7] El poco uso del dominio es en España se debe a lo costoso que es en comparación con los dominios genéricos y a las restricciones que posee la inscripción de los dominios.

nios com, org y net en Chile son muy poco usados, no superando el 0,5 % en los sitios ni el 2 % en las páginas.

Chile			España		
Enlaces	Dominio	Tipo	Enlaces	Dominio	Tipo
24610	boonic	C(S)	843	adobe	C
6016	uchile	E	595	boe	G
4648	vivastreet	C	520	elmundo	C
2402	olx	C	518	mec	G
1773	puc	E	503	elpais	C
1496	terra	C	500	terra	C
1450	portalciudadano	C	452	csic	G
1307	123	C	448	gva	G
1197	sii	G	400	abc	C
1171	gov	G	394	mtas	G

Cuadro 10: *Dominios con más enlaces a ellos*
C: Comercial, E: Educacional, G: Gobierno.

	Chile (Sitios)	Chile (Páginas)	Espña(Sitios)	España(Páginas)
local (cl/es)	99.62 %	98.12 %	15.965 %	56.033 %
com	0.29 %	1.59 %	65.026 %	31.436 %
org	0.04 %	0.05 %	7.581 %	5.950 %
net	0.04 %	0.24 %	7.387 %	4.954 %

Cuadro 11: *Distribución de los tipos de dominios.*

Chile		España	
Dominio	Sitios	Dominio	Sitios
com	61.42 %	com	49.99 %
org	13.23 %	org	8.69 %
net	7.08 %	net	6.07 %
ar	3.62 %	tk	3.25 %
info	2.25 %	de	3.13 %

Cuadro 12: *Distribución de los tipos de dominios.*

5.6. Dominios de primer nivel externos

Se han estudiando también los dominios a los que pertenecen los sitios internacionales enlazados desde las webs de Chile y España, es decir, sitios que pertenecen a webs nacionales de otros países. Se encontró una prevalencia de los enlaces a sitios de dominio com, org y net con proporciones similares en Chile y España. Los resultados se aprecian en la Tabla 12.
Es importante destacar que el número de enlaces externos e internos de las webs de Chile y España guardan una relación importante con el número de importaciones y exportaciones, lo que se puede ver con gran detalle en los estudios correspondientes a cada país o en [1].

6. Proveedores y Servidores

6.1. Software utilizado en los servidores

Cada servidor web puede entregar, si así lo desea, información respecto a las tecnologías que utiliza para generar las páginas, así como del software servidor y del sistema operativo que está ejecutando. En base a los servidores que entregan información, se determinó que en Chile el 66,7% de esos servidores ejecuta el software Apache, mientras que el 32,8% ejecuta Microsoft IIS. Otros servidores web no tienen una presencia notoria. En España la diferencia no es tan grande: la presencia de Apache llega al 46,89%; la de IIS, al 38,88%. Este margen menor tiene estrecha

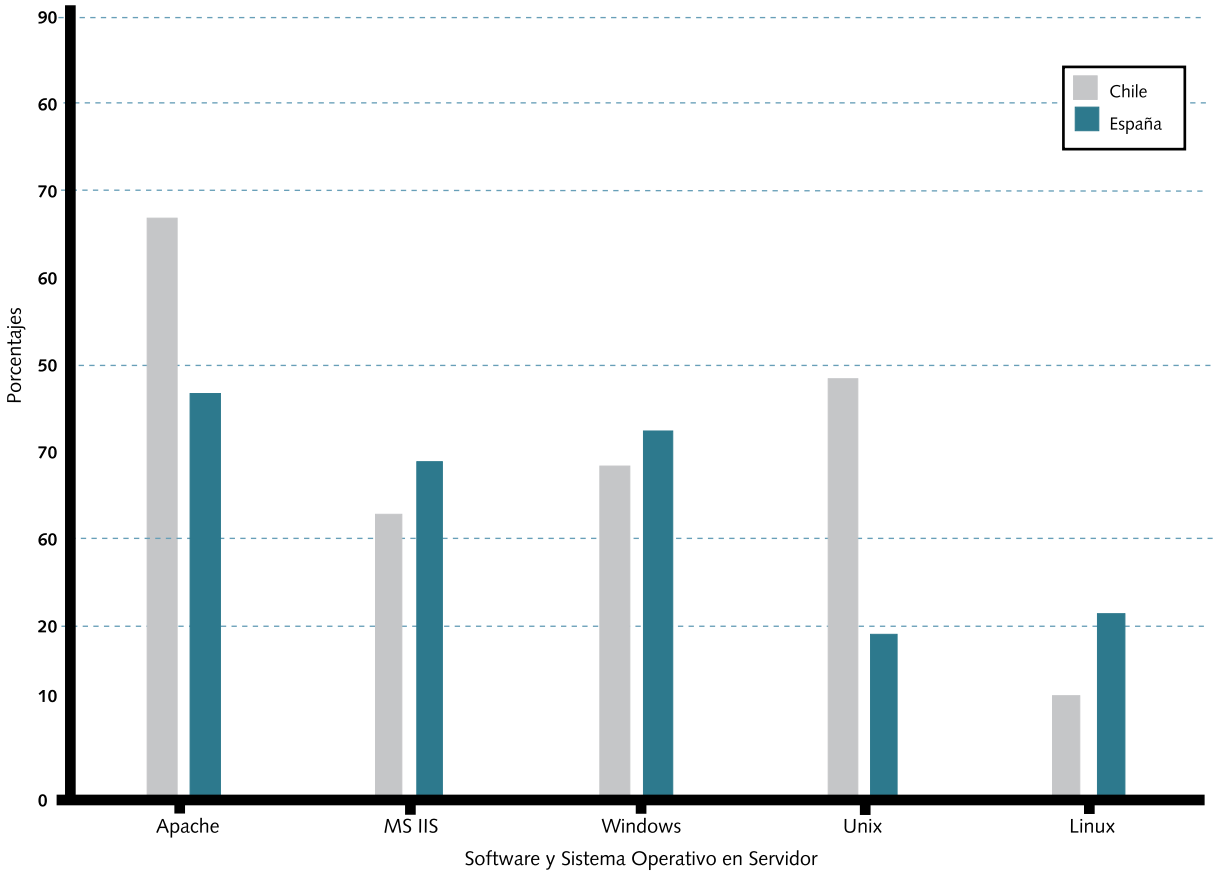


Figura 3: *Comparación entre software utilizado en los servidores.*

relación con el margen entre el uso de las tecnologías PHP y ASP.

Respecto a los sistemas operativos, en Chile cerca de un 60% utiliza servidores que corren bajo Linux/Unix, y un 38,5% de ellos ejecuta alguna versión de Microsoft Windows. En España la presencia de Windows, de un 42,59%, supera a la de Linux/Unix, que llega al 40,74%. Estas relaciones se pueden apreciar gráficamente en la Figura 3.

6.2. Proveedores y Números de IP

Mediante una búsqueda DNS Inverso determinamos los nombres de dominio de los proveedores con mayor presencia. En España son servidores dns, acens, terra, ono y veloxia; en Chile, ifxnw, virtuabyte, tchile y dattaweb.

En Chile encontramos cerca de 13.500 direcciones IP diferentes; en España, cerca de 24.000. En ambos casos la distribución de direcciones y dominios es muy sesgada: en Chile había 3 direcciones IP con más de 1.000 dominios, y en España 4 IPs hospedaban más de 1000 dominios. En cambio, las direcciones IP que sólo hospedaban un dominio son 8.391 en Chile y 16.565 en España. Tales sesgos se pueden representar mediante una ley de Zipf de parámetros 1,31 y 0,62 respectivamente.

7. Conclusiones

Este artículo se inició con la siguiente pregunta: ¿se puede hablar de una web nacional? Los estudios que se han comparado demuestran que sí, que si bien un sitio web típico existe sin conocer la existencia de los otros miles de sitios que componen su web nacional, cuando se juntan todos los sitios y se analizan se observan propiedades comunes, sin importar si el sitio lo creó un español, un chileno, un americano, un asiático o un europeo. Ciertamente las proporciones de sitios y documentos guardan una relación con los datos poblacionales del país al que pertenecen, al nivel de acceso a la red que exista en ese país, a su condición económica, y a un sinfín de otros factores; se podría concluir falsamente que las webs nacionales no comparten muchas propiedades. Se ha enseñado que efectivamente esto no es así; las webs de Chile y de España tienen similitudes bastante marcadas, en particular en el sesgo de las distribuciones de enlaces entre documentos, sitios y documentos, en sus tamaños y en sus vocabularios.

En este momento ambas webs deben ser muy diferentes a lo que eran al momento de ser recolectadas, ya que la web cambia y crece constantemente. Sin embargo, los diferentes estudios realizados han demostrado que a pesar de esta evolución sus propiedades se mantienen, y que incluso son similares a las propiedades globales. Se puede concluir que este tipo de comparación no sólo sirve para obtener datos estadísticos que estimen las propiedades de las colectas, sino
que también permiten conocer qué características permiten diferenciar a un documento de otro y así establecer una valorización para ellos, como bien puede ser PageRank.

Por ejemplo, en la actualidad los buscadores no solamente ordenan los resultados de búsqueda de acuerdo al PageRank de un documento, también consideran el lugar donde se produjo el calce entre las palabras de la consulta y el contenido del documento. De acuerdo a lo dicho en los estudios, los títulos de los documentos no sólo son importantes por el hecho de ser encabezados de las páginas, sino que también lo son por ser, de cierto modo, escasos, por lo que son bastante valorados. Cuando un webmaster realiza este tipo de consideraciones para su sitio, se dice que está utilizando técnicas SEO (*Search Engine Optimization*), es decir, emplea diferentes métodos para que un documento sea más valorado que otros por los recolectores de los motores de búsqueda.

Otro punto importante son las anomalías que presenta el SPAM y las nuevas tecnologías en la generación de páginas dinámicas (como *url rewriting*, es decir, entregar parámetros en la URL). La única forma de detectarlas es conocer bien su funcionamiento y presentación, y la única forma de determinar bien esos factores es analizando la mayor cantidad posible de documentos en busca de patrones o heurísticas que

permitan identificarlos después, con el fin de conseguir colectas menos distorsionadas.

Finalmente, cada una de estas colectas, de las que se han hecho en el pasado y de las que se harán en el futuro, es una fotografía del estado de la web. Ciertamente estas fotografías nunca cesarán en el futuro, porque la web ha llegado para cambiar la vida de todos y ser un nuevo medio de comunicación y expresión. Las referencias indicadas permitirán al lector interesado adentrarse en estos estudios y en sus relaciones con la informática, la web global y la sociedad.

Agradecimientos

Agradecemos los excelentes comentarios de Marcelo Garrido que ayudaron a mejorar sustancialmente este trabajo.

Referencias

[1] Ricardo Baeza-Yates and Carlos Castillo. Relationship between web links and trade. *In WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 927–928, New York, NY, USA, 2006. ACM Press.

[2] Ricardo Baeza-Yates and Carlos Castillo. *Crawling the infinite web.journal of Web Engineering,* 6(1):49–72, February 2007.

[3] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2), May 2007.

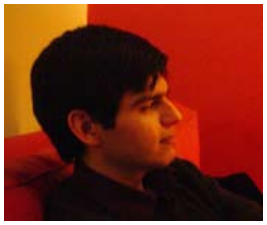
[4] Ricardo Baeza-Yates, Carlos Castillo, and Eduardo Graells. Características de la Web Chilena 2006. Technical report, Center for Web Research, University of Chile, 2007.

[5] Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. Characteristics of the Web of Spain. *Cybermetrics*, 9(1), 2005.

[6] A.L. Barabási and RE Crandall. Linked: The New Science of Networks. *American journal of Physics*, 71:409, 2003.

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

Ricardo Baeza-Yates es VP de Investigación para Europe y Latinoamérica, liderando los laboratorios de Yahoo! Research en Barcelona, España y Santiago, Chile. Hasta 2005 fue director del Centro de Investigación de la Web en el Departamento de Ciencias de la Computación de la Escuela de Ingeniería de la Universidad de Chile; y catedrático ICREA en el Dept. de Tecnología de la Universitat Pompeu Fabra en Barcelona, España. Mantiene vínculos con ambas universidades como profesor jornada parcial.



Eduardo Graells. De 24 años, asiduo a la lectura y los videojuegos, es alumno memorista de Ingeniería Civil en Computación en la Universidad de Chile. Sus áreas de interés son Recuperación de la Información, Minería Web y Computación Gráfica. Ve (quiere ver) en la web actual la biblioteca de Babel con la que tanto soñó Jorge Luis Borges.