Caracterizando la Web Chilena*

Ricardo Baeza-Yates *Carlos Castillo*†
Depto. de Ciencias de la Computación
Universidad de Chile
Blanco Encalada 2120, Santiago, Chile

Resumen

Entre mayo y junio del año 2000 se llevó a cabo un estudio sobre las características de la Web Chilena, basado en datos obtenidos con el recolector de páginas del buscador TodoCL. Dicho estudio contempla tanto características individuales de las páginas, como del conjunto de las páginas a nivel de sitio y de dominio. Se presentan numerosos datos estadísticos y modelos que configuran los aspectos fundamentales de la Web Chilena.

Mediante una clasificación de los dominios se muestra una representación concisa de la conectividad entre ellos, así como características de las estructuras que esta representación revela.[1]

Keywords: WWW, Análisis de Links, Estadísticas, Web Chilena.

1 Introducción

El interés de realizar una descripción de la Web es proveer de información para aplicaciones técnicas, comerciales y sociológicas, en particular para minería de datos y de comportamiento de los usuarios. La Web es altamente dinámica y su caracterización permite entender como evoluciona. Además es importante estudiar subconjuntos de la Web, delimitados por criterios culturales o geográficos: para este estudio, usamos el buscador TodoCL (`http://www.todocl.cl/`), desarrollado en el Departamento de Ciencias de la Computación de la Univ. de Chile.

En el año 1993 se instaló el primer servidor Web Chileno (y uno de los primeros en Iberoamérica) en el DCC de la Universidad de Chile. Conforme aparecieron más servidores, se decidió implementar un mapa sensible de Chile [2] dividido por regiones, en el cual aparecería cada sitio basado en su localización geográfica. Este mapa permitía visualizar el estado de la Web Chilena de una forma clara e inequívoca, pero 6 meses más tarde era imposible seguir manteniéndolo, pues diariamente aparecían nuevos sitios. Hoy es necesaria una nueva forma de visualización de los sitios, basada en una estructura virtual, que utilice criterios más complejos que el geográfico.

Con el boom de la Web, en los últimos años se ha comenzado a entender cómo es ella, desde el punto de vista de su estructura y de cómo se usa. En particular, un estudio reciente muestra el hecho de que pocos sitios web concentran la mayoría de las visitas de los usuarios [4]. Con respecto a la macroestructura, en [2] se analiza una posible caracterización del Web global basada en la conectividad y se propone la clasificación de sitios que adoptamos en este trabajo. En [5] se estudia una caracterización de las páginas que utiliza información tanto de conectividad como del comportamiento de los usuarios que utilizan un sitio.

Dado el tamaño y crecimiento de la Web, estos estudios son difíciles de realizar periódicamente. Sin embargo, esto es más sencillo en subconjuntos de ella, lo que permite verificar si las características globales

*Trabajo parcialmente financiado por Proyecto Fondecyt 990627.

†{rbaeza,ccastill}@dcc.uchile.cl

[1] Este artículo resume algunos de los resultados más destacados. El estudio completo se encuentra disponible en `http://www.todocl.cl/stats.phtml`.

[2] `http://sunsite.dcc.uchile.cl/chile`

se replican a estructuras locales y su nivel de desarrollo. Por ejemplo, una descripción al nivel de página y sitio fue realizada durante 1999 para la Web Brasileña [3] usando **TodoBR** (`http://www.todobr.com.br`). TodoBR es un buscador de páginas desarrollado en el Departamento de Ciencia de la Computación de la Universidad Federal de Minas Gerais, en Belo Horizonte, Brasil. Del mismo modo, el estudio de la Web Chilena es interesante por si mismo, y permite además su comparación con estudios similares.

Una motivación comercial importante es que muchos negocios en Internet están mantenidos por medio de publicidad, en donde el beneficio de cada sitio comercial depende directamente del número de visitas que recibe el sitio. Este número de visitas está fuertemente correlacionado con la cantidad de referencias que tiene un sitio Web en el resto de la colección. Además las visitas siguen una ley del "ganador se lo lleva todo"pues sólo unos pocos sitios atraen prácticamente toda la atención de los usuarios.

En este artículo realizamos un estudio similar a los anteriores, pero el análisis de conectividad lo realizamos en base a dominios y no a páginas, ya que creemos que la conectividad a nivel macro es más interesante que a nivel micro. Además, realizamos el primer estudio de la correlación que existe entre la macroestructura de un subconjunto de la Web con páginas clasificadas manualmente y el comportamiento de los usuarios cuando buscan. En la siguiente sección explicamos brevemente la metodología del estudio. En las restantes secciones caracterizamos la Web Chilena a nivel de colección, páginas, sitios y dominios, respectivamente. Finalizamos con las conclusiones principales de nuestro estudio y proponemos extensiones al mismo.

2 Conceptos Básicos

Un buscador Web que utilice indexación automática, como TodoCL - al igual que Altavista, AllTheWeb, Inktomi o NorthernLight - usualmente incluye un recolector de páginas o *spider* que comienza recorriendo e indexando un conjunto de sitios predeterminado (puntos de partida), para luego seguir indexando las páginas que son apuntadas desde estos sitios mediante un procedimiento recursivo. Este proceso es realizado simultáneamente por varios spiders a la vez, los que se comunican con un planificador o *scheduler* de indexación. Dicho proceso puede ser optimizado si se conocen a priori algunos datos sobre las páginas que integran la colección: el tamaño de éstas, su frecuencia de actualización y el número promedio de links que posee cada una.

Para obtener estos datos, se utilizó el recolector y el *scheduler* de visita a sitios Web de TodoCL, que son adaptaciones del recolector desarrollado para TodoBR[7]. Se consideraron fundamentalmente páginas bajo el dominio `.CL` más algunas páginas en el dominio `.NET` pertenecientes a empresas chilenas, principalmente proveedores de acceso a Internet (ISPs).

Este recolector realiza un proceso de filtrado de las páginas, en el cual ellas son convertidas a formato de texto plano. Se usan filtros para varios formatos comunes de documento, incluyendo `HTML`, `PDF`, `PostScript` y `Word`. Los archivos binarios (gráficos, archivos comprimidos u otros) no se incorporan a la colección. Se utilizó una heurística para eliminar bloques de archivo que hubieran pasado a través de los filtros sin ser documentos de texto, con lo que se descartó cerca del 4% de la colección, en su mayoría archivos binarios con encabezados malformados.

La descripción que presentamos se divide en 4 niveles:

- Colección: cifras globales y estudio del vocabulario.
- Página: tamaño, tipo de documento, edad e idioma.
- Sitio: profundidad de las páginas, número de páginas por sitio, contenido de texto por sitio.
- Dominio: número de referencias hacia y desde un dominio, representación de la estructura global de hipervínculos entre dominios, preferencias de los usuarios.

3 Nivel Colección

3.1 Cifras Globales

En la tabla 1 se muestra el tamaño de la base para el estudio, obtenida en 15 días de recolección y que estimamos corresponde a más del 95% de la Web Chilena.

Puntos de partida	19.390
Páginas	730.673
Sitios	10,352
Dominios	9,102
Tamaño de la Colección	2.3 Gb
Tamaño del Vocabulario	1.9 Millones

Tabla 1: Tamaño de la colección.

Los puntos de partida corresponden a 19.200 nombres de dominio y alrededor de 200 direcciones ingresadas por los usuarios de TodoCL. El tamaño de la colección considera *sólo el texto* de los archivos que fue posible convertir.

La mayoría de los dominios (90%) se encuentra en Santiago, lo siguen las regiones V (780 dominios), VIII (268 dominios) y X (211 dominios).

3.2 Vocabulario

El conjunto de palabras distintas en la colección se denomina *vocabulario*. El modelo más utilizado para el tamaño del vocabulario (V) en función del tamaño de la colección (n) es la *Ley de Heaps* [8], que establece que:

$$V = Kn^{\beta}$$

donde los parámetros K y β dependen de la colección.

Se estimó $K \approx 2.22$ y $\beta \approx 0.63$ para la colección del texto encontrado en páginas de la Web Chilena. El modelo sublineal se cumple con bastante precisión, y muestra que la cantidad de palabras tiende a ser más grande que en las colecciones de texto editado en inglés donde $\beta \approx 0.5$. Esto se debe en parte a la cantidad de palabras con errores y a la diversidad de lenguajes.

3.3 Palabras más Frecuentes

Descartando los números, así como los artículos, las preposiciones y otras palabras funcionales, las palabras que aparecen en más documentos en la web Chilena son: "Chile", "CL", "Home", "Información", "Copyright", "Santiago", "Mail", "Internet", "WWW", "Software"y "Page". Es interesante notar que el 31% del texto de las páginas Chilenas incluye la palabra "Chile".

4 Nivel Página

4.1 Tamaño

Se estudió la distribución del tamaño de cada página y de la porción de este tamaño que corresponde a texto, descartando ilustraciones y comandos de formato en los tipos de datos analizados.

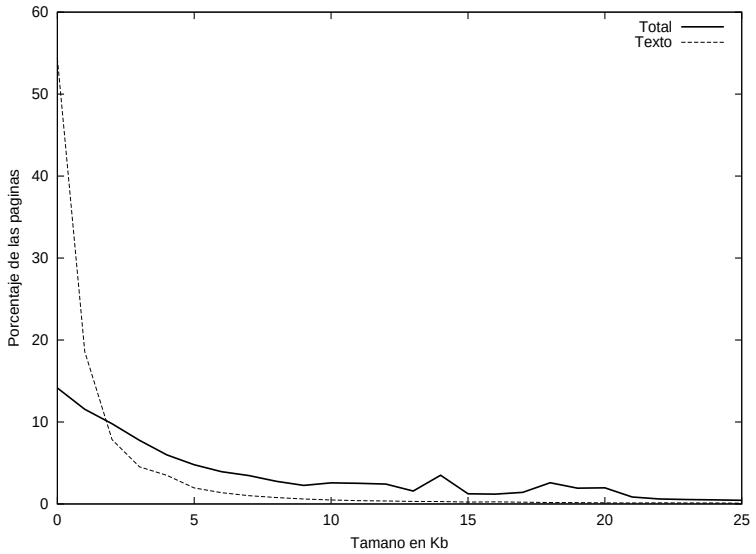

Figura 1: Tamaños de Texto en Páginas.

El gráfico con los tamaños se muestra en la figura 1. La mayoría de las páginas tienen poco texto (incluso un 50% aproximadamente sólo tiene imágenes o tags), y el promedio de texto es de 3.4 Kb, mientras que para la página en su totalidad es de 15.3Kb.

En promedio el 75% del tamaño de un archivo HTML es usado por las marcas de formato (*tags*), siendo texto sólo el 25% restante. Los archivos mayores de 40kb tienen algo más de texto, alrededor de un 30% del tamaño total del archivo.

Sólo un porcentaje muy pequeño (3% de las páginas) tiene más de 40Kb.[3]

4.2 Tipo

El tipo de documento más común es HTML, con más del 95% de las páginas, los demás formatos de documento lo siguen bastante más atrás, destacandose Texto Plano (2.14%), Adobe PDF (0.85%) y Microsoft Word (0.75%). Estos datos no se obtuvieron basándose en la extensión del archivo, sino en el resultado de los filtros y de los primeros bytes de cada documento (denominados "*magic numbers*").

4.3 Edad

Entendemos por *edad* de una página el tiempo transcurrido entre su última actualización y el momento en que el recolector la visita. La mayoría de los servidores Web entregan junto con la página la fecha en que la página fue modificada por última vez, pues esto es parte del estándar HTTP; sin embargo, una cantidad no despreciable de las páginas (19%) no tienen asociada una fecha de última modificación, o tienen una fecha incorrecta[4].

La edad de las páginas sigue una distribución potencial, consistente con las observaciones para la Web global en [6]. Un 2.6% de las páginas se reporta con menos de 1 día de antigüedad; esta es una cantidad bastante considerable (más de 10000 páginas) pero incluye las que son generadas dinámicamente en el

[3]Por motivos de eficiencia, el buscador trunca las páginas de más de 1Mb, por lo que no se puede determinar el tamaño máximo de texto en una página. Los archivos con más texto usualmente se distribuyen en formato `PDF`.

[4]En algunos casos, varios años en el futuro.

momento de ser accesadas. Un resumen acumulativo con la última actualización de las páginas se puede ver en la tabla 2. La edad promedio es de 274 días y la mediana 145 días, de modo que aproximadamente la mitad fue actualizada hace menos de 5 meses.

Hoy	2.6%
Esta semana	8.4%
Este mes	20.9%
Este semestre	54.1%
Este año	75.9%
Últimos 2 años	89.5%
Últimos 3 años	96.4%

Tabla 2: Última actualización de las páginas.

4.4 Idioma

Observaciones preliminares sobre muestras de 200-300 documentos indicaron que además del español, un porcentaje importante de las páginas Chilenas están escritas en inglés. Las mismas observaciones hacen suponer que entre un 1-2% de las páginas están en ambos idiomas a la vez. Dado que otros idiomas como el francés, portugués y alemán fueron observados, pero sumados no alcanzan a representar el 1% del total de documentos, nos concentramos en la tarea de determinar cuál es el porcentaje de documentos escritos en inglés sobre el total de documentos.

Para determinar esta cifra, partimos desde el hecho conocido de que aproximadamente un 40%-50% de las palabras en un texto son *stopwords* [1] o palabras funcionales, lo que es común a ambos idiomas. Esto permite utilizar una heurística de discriminación de lenguaje basada en el hecho de que un texto normal [5] en inglés contiene muchas stopwords de inglés y un texto normal en español contiene muchas stopwords en español. Utilizando esta heurística, se obtiene que alrededor de un 7-8% de las páginas de la Web Chilena están escritas en inglés.

4.5 Multimedios y Otros Formatos

El 62% de los archivos HTML revisados contenía al menos una imágen gráfica, contando *bullets* y otros tipo de ícono. De los archivos con imágenes, un 61% utiliza sólo formato `GIF` (*Compuserve Graphics Interchange Format*) y un 11% sólo `JPEG` (*Joint Photographics Expert Group*) - un 27% mezcla imágenes de ambos formatos en la misma página. El uso del formato `PNG` (*Portable Network Graphics*) fue observado sólo a nivel muy incipiente (menos del 1%).

Respecto a otros tipos de archivo, las páginas que incluyen archivos comprimidos y/o programas alcanzan aproximadamente un 4% del total, mientras que las que incluyen archivos de audio y/o video menos del 1%.

5 Nivel Sitio

5.1 Número de Páginas

El 52% de los sitios en la colección tienen sólo una página y prácticamente todos los sitios tienen menos de 100 páginas, como se puede ver en la figura 2. De los 20.000 dominios registrados sólo se utilizan 10.000

[5] Quedan fuera casos anómalos como por listas de nombres propios, que no tienen stopwords.

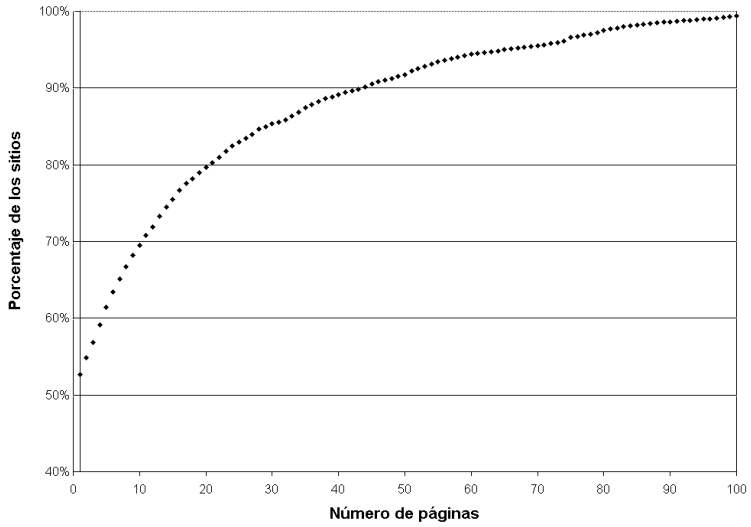

Figura 2: Número de páginas por sitio

para poner sitios web, y de estos sitios sólo 5.000 vayan más allá de una simple portada, lo que habla sobre la tendencia a "estar en Internet"de las empresas y organizaciones más que a "hacer cosas en Internet".

Las páginas están muy concentradas en unos pocos sitios, por ejemplo, la mitad de las páginas de la colección se encuentran en los 1300 sitios más grandes, cada uno con 70 o más páginas. Otro indicio de este fenómeno es que los 100 sitios más grandes (1000 o más páginas) contienen un tercio de las páginas de la Web Chilena.

5.2 Profundidad de las Páginas

Si bien entre el conjunto de los sitios las relaciones por hipervínculo tienen una relación "en red", la estructura de las relaciones al interior de un sitio resulta más bien jerárquica, con una portada de la cual se cuelgan varias secciones y subsecciones. Una forma de estudiar este árbol jerárquico de relaciones es observar la *profundidad* de las páginas dentro del árbol.

Una aproximación razonable a la profundidad de una página dentro del árbol jerárquico de relaciones es su profundidad física dentro del árbol de directorios. Así, por ejemplo, una página cuya URL es: `www.uchile.cl/aa.html` tiene profundidad 1, `www.uchile.cl/bb/aa.html` tiene profundidad dos y así sucesivamente. En términos gruesos se observa que la mitad de las páginas está a profundidad 2 o 3. La otra mitad se encuentra distribuída entre las profundidades 4, 5 y 6, en orden decreciente de importancia.

5.3 Tamaño Total

El tamaño del texto en cada sitio (la suma del tamaño del texto de las páginas que lo componen), sigue una distribución similar al número de páginas por sitio. El 1% de los sitios más grandes en contenido, aportan el 60% del texto total en la colección. En el otro extremo, los sitios de una sóla página (como se mostró más arriba, cerca del 50% de los sitios) prácticamente no aportan texto. Esto se muestra en la figura 3, en la que también se incluye la distribución del tamaño total de las páginas web, considerando *tags*.

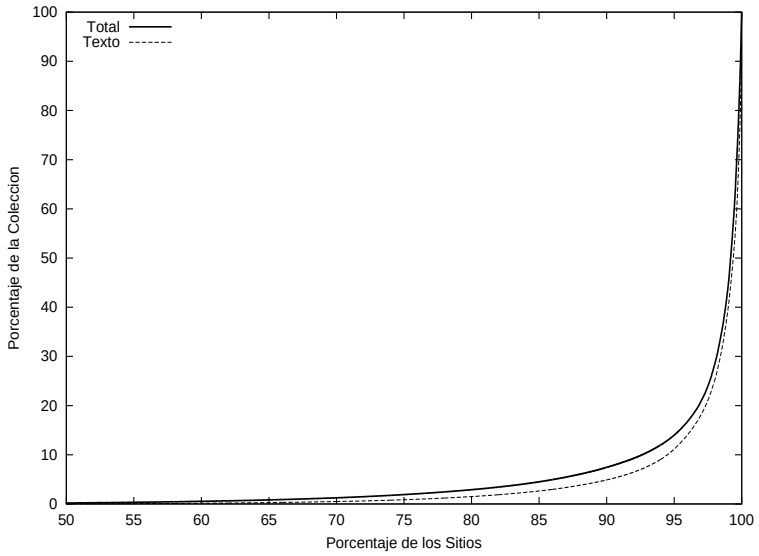

Figura 3: Tamaño del texto y tamaño total de las páginas, agrupadas por sitio

6 Nivel Dominio

Se estima que las páginas bajo un mismo dominio tienen relación entre sí [6]. Las relaciones entre dominios pueden representarse como un grafo dirigido, en que cada vértice representa un dominio D_i y un arco va de D_i a D_j si existe un link desde una página en el primer dominio hacia una página en el segundo (es decir, varios enlaces se colapsan a uno sólo). En adelante llamaremos a este grafo el *grafo de links* (enlaces), y mostraremos un análisis del mismo sobre una muestra de aproximadamente 6200 dominios.

6.1 Grado Interno y Externo

El número de enlaces externos[7] hacia un dominio (grado interno en el grafo de links) y el número de enlaces externos desde un dominio (grado externo), siguen una distribución potencial de exponente negativo, de acuerdo a la ecuación

$$frecuencia = \alpha(Grado)^{-\theta} .$$

Usando esta ecuación, estimamos para el grado interno, $\alpha = 1781,9$ y $\theta = 1,4092$ y para el grado externo, $\alpha = 1265,3$ y $\theta = 0,9744$.

Los 10 dominios chilenos hacia los que llegan más enlaces corresponden a los sitios de la Universidad de Chile, Chilnet, El Mercurio, La Brújula, la Universidad Católica, la Dirección Meteorológica de Chile, La Tercera, el Banco Central, la Universidad de Concepción y el Servicio de Impuestos Internos

6.2 Largo de Caminos

Nos preguntamos si dados dos nodos D_1 y D_2 (escogidos al azar del grafo de links) existe un camino dirigido desde D_1 hasta D_2, y si es así, cuál es el número máximo y el promedio de dominios que habría que visitar.

[6] Aunque es usual que se utilicen varios subdominios para usos distintos y que un mismo sitio tenga páginas no relacionadas, en general las páginas bajo el mismo dominio están relacionadas.

[7] Es decir, descontando los enlaces dentro del mismo dominio.

Si no consideramos la dirección de los arcos, casi siempre existe un camino, pues la componente conexa más grande ocupa el 94% de los dominios.

Considerando la dirección de los links, sólo un 25% de los nodos pertenece a la componente *fuertemente* conexa principal. Sólo dentro de esta componente es posible encontrar un camino entre dos nodos cualesquiera. Estudios sobre esta porción del grafo muestran que el camino promedio pasa por 3 dominios y tiene un largo máximo de 13 dominios (este es el diámetro de la componente conexa).

6.3 Macroestructura

En [2] se propone una forma de clasificar las páginas Web en base a su conectividad. Se comienza observando la presencia de una gran componente fuertemente conexa [8] en el grafo de links, que incluye aproximadamente un 25% de las páginas. Adoptamos la siguiente nomenclatura para clasificar los dominios de acuerdo a su relación con esta componente fuertemente conexa principal:

- `MAIN`: Componente fuertemente conexa principal (23%).
- `IN`: Dominios desde los cuales `MAIN` es alcanzable (tienen enlaces hacia `MAIN`, o tienen enlaces hacia sitios que apuntan a `MAIN`, y así sucesivamente) (14%).
- `OUT`: Dominios que son alcanzables *desde* `MAIN` (45%).
- `TENTÁCULOS`: Dominios que tienen relación con `IN` o `OUT`, pero no con `MAIN` (14%).
- `ISLAS`: Dominios que no tienen relación con ninguna de las anteriores (4%).

Un dominio pertenece a una y sólo una de las anteriores. Se estudiaron también subconjuntos de estas componentes, denominados de la siguiente forma:

- `MAIN-IN`: Dominios en `MAIN` que tienen enlaces directos desde `IN` (5%).
- `MAIN-OUT`: Dominios en `MAIN` que tienen enlaces directos hacia `OUT` (8%).
- `MAIN-MAIN`: Intersección de las dos anteriores (2%).
- `MAIN-NORM`: Dominios en `MAIN` que no están en `MAIN-IN` ni en `MAIN-OUT`, es decir, que no tienen enlaces directos desde o hacia sitios fuera de `MAIN` (11%).
- `TENTÁCULOS-IN`: Sitios relacionados con `IN`, pero no con `MAIN` (3%).
- `TENTÁCULOS-OUT`: Sitios relacionados con `OUT`, pero no con `MAIN` (9%).
- `TÚNEL`: Dominios que permiten conectar a `IN` y `OUT` sin pasar por la componente principal, corresponden a una clase especial de `TENTÁCULOS` (1%).

Las relaciones de conectividad, así como el grado interno y externo promedio de cada componente se observan en la figura 4 que corresponde a una representación esquemática de la Web Chilena.

Los sitios en la componente `IN` se identifican con sitios nuevos que poseen referencias hacia la componente principal, pero que no poseen una referencia recíproca desde aquella componente (por ejemplo: no pertenecen a ningún directorio todavía), mientras que los de la componente `OUT` son en su mayoría sitios corporativos que proveen de información sobre alguna organización sin poner enlaces hacia otro dominio. También pueden representar páginas de mayor antigüedad, creadas antes que la mayoría de las páginas en MAIN y que no pasaron a formar parte del núcleo de páginas más conocidas.

También se observó que los nodos que no están en la componente principal prácticamente no se enlazan entre sí, lo que concuerda con las medidas de número de links hacia y desde los dominios en cada componente.

[8] En el caso chileno, la segunda componente fuertemente conexa tiene sólo 10 sitios.

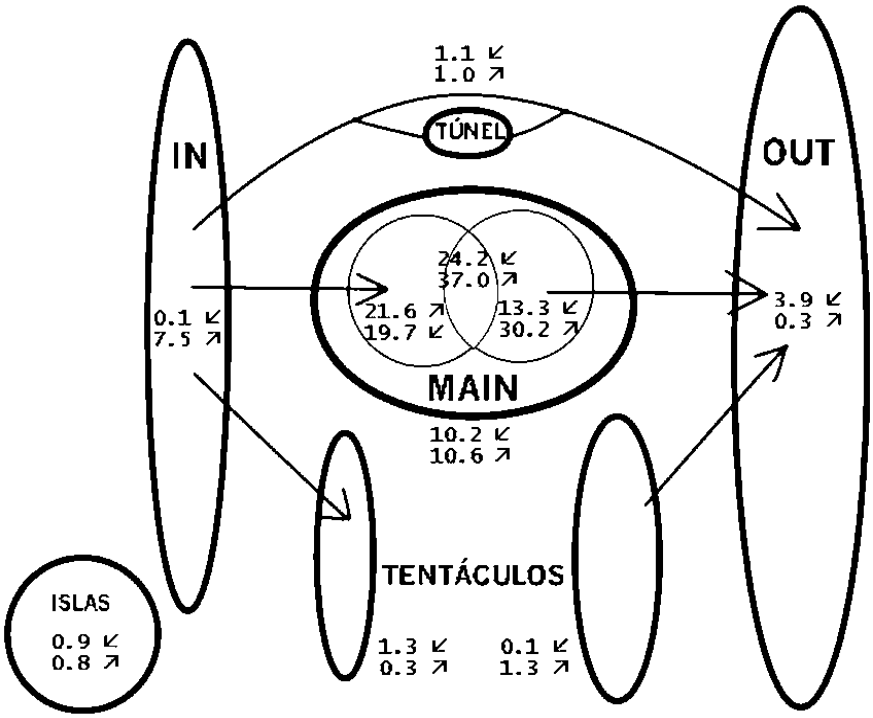

Figura 4: Macroestructura de hipervínculos, con grados interno y externo promedio.

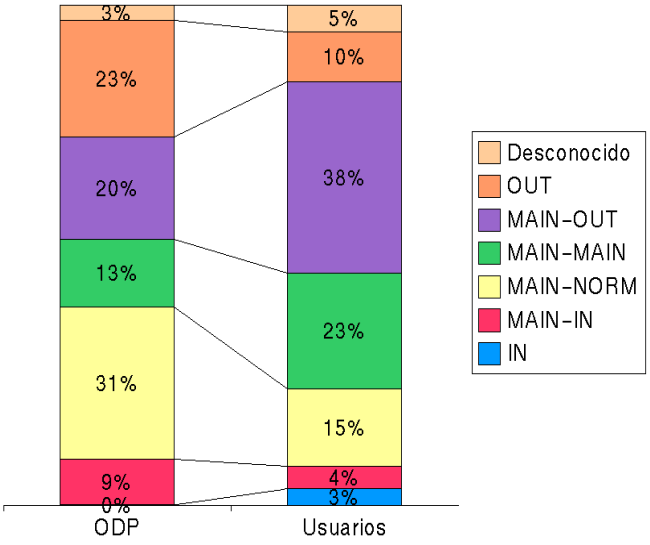

Figura 5: Ubicación de los sitios escogidos.

6.4 Preferencias de los Usuarios

El objetivo de esta sección es estudiar a qué componentes pertenecen mayoritariamente los sitios escogidos por usuarios, y tener indicios de si existen o no diferencias sustanciales entre los sitios a los que un usuario accede si utiliza una máquina de búsqueda automatizada o un directorio con sitios clasificados a mano.

Se utilizaron dos muestras independientes: Editores ODP y Usuarios TodoCL. La muestra ODP contempla 3.100 sitios clasificados por editores del Open Directory Project [9] en la categoría *World/Español/-Regional/Chile* que corresponden a 1.000 dominios dinstintos bajo `.cl`.

La muestra Usuarios TodoCL corresponde a la observación de 18.000 enlaces seguidos por los usuarios de entre los contenidos en las páginas de respuesta, que pertenecen a 2.500 dominios distintos. TodoCL cuenta con un sistema de redireccionamiento que permite tener un registro de los enlaces escogidos para cada consulta. Ambas muestras pueden interpretarse como: "Los sitios de la web Chilena que cumplen ciertos criterios de calidad"(ODP) y "Los sitios de la web Chilena que parecen relevantes a los usuarios al ser entregados por la máquina de búsqueda"(Usuarios).

La ubicación de los sitios escogidos por editores ODP y usuarios TodoCL en las componentes antes descritas da origen a la figura 5. La primera observación es el hecho de que la mayoría de los sitios escogidos por editores ODP se encuentran en `MAIN-NORM`[10], mientras que la máquina de búsqueda usada tiende a llevar a los usuarios hacia páginas que son usualmente directorios de otras páginas (`MAIN-OUT`), por el hecho de que éstas incluyen a menudo muchas palabras distintas y eso las lleva a aparecer como respuesta en varias consultas diferentes.

La segunda observación es que el número de sitios en la componente `OUT` ofrecidos por TodoCL es más bien bajo y probablemente aumente con el uso de algún algoritmo de análisis de enlaces como PageRank [9].

[9] Disponible en `http://odp.org`. Este es un proyecto que entrega su base de datos bajo licencia *Netscape Public License*, una variante de GPL, que es utilizado como directorio de páginas en TodoCL.

[10] Esto es, dominios que están en `MAIN`, pero que no tienen enlaces hacia o desde dominios fuera de `MAIN`

7 Conclusiones

Se verificó que varios de los resultados establecidos para la red global eran también válidos para la Web Chilena, y encontramos correspondencias con el estudio realizado en la Web Brasileña. Esto indicaría que tal como otros fenómenos de Internet, la estructura de la Web es altamente autosimilar (es decir, la estructura no se modifica ante cambios de escala).

Las características locales más destacadas son que la tasa de utilización de los dominios inscritos es aproximadamente de un 50%, y la gran mayoría de los sitios existentes cuentan con sólo una página (estas son las llamadas "páginas de presencia en Internet", que contienen usualmente una foto, algunos párrafos de texto y la dirección de e-mail de la empresa). A pesar que existen muchas páginas, éstas en general presentan poco contenido y están concentradas en unos pocos sitios.

Se extendieron los resultados de [2], estudiando características internas de las estructuras observadas en el grafo de links, en particular el hecho de que fuera de la componente fuertemente conexa principal `MAIN` la conectividad es bastante baja.

Dicho de otra forma, fuera del 25% de los sitios que forman la componente principal, la característica de *red* que se menciona profusamente (tanto en la literatura especializada como en la de difusión) no existe, presentándose mayoritariamente sitios desconectados entre sí. Esto concuerda con los modelos sobre el número de enlaces desde y hacia cada página, que muestran que la mayoría de los links van hacia y desde un conjunto más bien pequeño de páginas.

Finalmente, se mostraron características cualitativas de las componentes utilizando datos provistos por usuarios, en particular la importancia relativa de cada componente en sus preferencias. Se encontraron diferencias importantes entre los sitios a los que accede un usuario que utiliza una máquina de búsqueda por palabras clave como TodoCL o un directorio jerárquico como ODP.

Referencias

[1] R.Baeza-Yates,B.Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley-Longman 1999.

[2] A. Broder, R. Kimar, F. Maghoul, P. Raghavan, S. Rajagopalan, R.Stata, A. Tomkins, J. Wiener, *Graph Structure in the Web*, Proc. of WWW9, Mayo 2000. `http://www.almaden.ibm.com/cs/people/pragh/www9.html`.

[3] Eveline A. Veloso, E. de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender, B. Ribeiro-Neto, N. Ziviani, *Um Retrato da Web Brasileira*, Simposio Brasileiro de Computacao, Curitiba, Brasil, Julio 2000.

[4] Adamic and Huberman, *The nature of markets on the World Wide Web*, Xerox PARC Technical Report, 1999, `http://www.parc.xerox.com/istl/groups/iea/www/webmarkets.html`

[5] P. Pirolli, J. Pitkow, R. Rao, *Silk from a Sow's Ear: Extracting Usable Structures from the Web*, `http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html`

[6] B. Brewington, G. Cybenko, *How Dynamic is the Web ?*, Proc. of WWW9, Mayo 2000.

[7] A. da Silva, E. Veloso, P. Golgher, B. Ribeiro, A. Laender, N. Ziviani, *CoBWeb: A Crawler for the Brazilian Web.* en *Proc. of the* 6^{th} *International Symposium on String Processing and Information Retrieval* (SPIRE '99) páginas 184-192. Carleton University Press, 1999.

[8] J. Heaps, *Information Retrieval - Computational and Theoretical Aspects.* Academic Press, 1978.

[9] L. Page, S. Brin, R. Motwani and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford University, 1998.