Caracterización del Espacio Web de Argentina

Gabriel H. Tolosa
Universidad Nacional de Luján, DCB – Laboratorio de Redes de Datos
tolosoft@unlu.edu.ar

Fernando R. A. Bordignon
Universidad Nacional de Luján, DCB – Laboratorio de Redes de Datos
bordi@unlu.edu.ar

Ricardo Baeza-Yates
Universidad de Chile, DCC, Centro de Investigación de la Web (CIW)
Yahoo! Research Latin America
ricardo@baeza.cl

Carlos Castillo
Yahoo! Research Barcelona
chato@yahoo-inc.com

Abstract

This article presents the results of research on the characterization of the Argentinian web domain over a sample of almost 10 million web pages from 150.000 sites collected in the early 2006. Particularly, we have studied page contents, link structure and technologies used in the construction of the sites. The results are consistent with earlier research on other national web domains. This study reveals a number of interesting facts: To begin with, there is a significant proportion (97.6%) of ".com.ar" domains. As regards page contents, we have found a predominance of terms related to commercial activity. However, terms found in site names, extracted from their URLs, are mostly related to tourism. 72% of the pages have been created or modified in the last year, which indicates that the Argentinian web space is growing quickly. As for technologies, 48% of the pages from the sample are static and 52% dynamic, the latter being mostly built using free tools. Besides, 76% of the sites are hosted in servers geographically located in Argentina. These two facts show there is an important web-related technological development and communication infrastructure in Argentina.

Keywords: Web characterization, Argentinian National Domain, Web Measurement, Link Analysis

Resumen

En este trabajo de investigación se caracteriza el espacio web argentino a partir del análisis de una muestra, tomada a principios del año 2006, cercana a los 10 millones de páginas extraídas de 150.000 sitios. En particular, se realizó análisis de contenidos, de enlaces y de tecnologías utilizadas para construir sitios. Los resultados obtenidos son consistentes con los de otros espacios webs nacionales. Del estudio surgen las siguientes observaciones: Existe una importante proporción de dominios "com.ar" (97.6%). En lo referente al contenido, predominan términos relacionados con la actividad comercial, mientras que en los nombres de los sitios, aparecen mayormente términos relacionados con el turismo. El 72% de las páginas han sido creadas o modificadas en el último año, lo cual indica que el espacio web argentino está creciendo aceleradamente. Con referencia a tecnologías, el 48% de las páginas de la muestra son estáticas y el 52%, dinámicas, las cuales se encuentran construidas en gran parte utilizando herramientas libres. El 76% de los sitios se hallan alojados en servidores que residen en Argentina. De los indicadores anteriores se desprende que existe un importante desarrollo tecnológico y de la infraestructura de comunicaciones de Argentina relacionada con la web.

Palabras clave: Caracterización de la Web, Dominio Nacional Argentino, Medidas de la Web, Análisis de Enlaces

1 Introducción

Actualmente, la World Wide Web es un espacio público utilizado por múltiples usuarios con objetivos diferentes. Originalmente, se presentaba como un repositorio distribuido que permitía compartir información y – aunque no ha perdido este objetivo – en la actualidad es un medio de publicación para diferentes usos como comercio, publicidad, educación, entretenimiento y contactos sociales, entre otros. Si bien la web se encuentra en constante crecimiento el estudio de características y tendencias entrega una valiosa información, tanto para entender su estructura como para desarrollar herramientas que faciliten la utilización de sus recursos. Algunos esfuerzos se han realizado para caracterizar el espacio web global. El World Wide Web Consortium realizó algunas de estas actividades hasta 1999 [23], mientras que el On Line Computer Library Center albergó un proyecto de esta naturaleza con información hasta el año 2002 [19].

El estudio de las características del espacio web es una tarea compleja que requiere de la utilización de recursos computacionales de gran escala debido a su tamaño y distribución geográfica. Debido a esto, se han realizado estudios a menos escala, específicamente de dominios nacionales [4, 5, 6, 15, 18], tomando muestras de dominios variados y utilizando diferentes estrategias de recolección. De acuerdo a [7], estas muestras presentan un buen balance entre diversidad y completitud, por lo que constituyen un conjunto de alto interés.

En este trabajo se presenta un estudio de caracterización del espacio web de Argentina, el cual presenta algunas particularidades interesantes que lo diferencian de otros. El estudio abarca las características principales reportadas en otros trabajos similares aunque aumentamos en nivel de análisis en algunos aspectos tratando de obtener nueva información. Para nuestro conocimiento, éste es el primer estudio sobre el espacio web de Argentina.

2 Características de la Web

La web puede ser modelada como un grafo dirigido (*webgraph*) donde los nodos corresponden a páginas HTML y los enlaces entre éstas son las aristas [11]. Formalmente, este grafo consiste en un conjunto de nodos, denotado como P y un conjunto A de aristas. Cada arista (denotada como $q \rightarrow p$) es un par ordenado (q, p) que representa un enlace o vínculo entre las páginas (nodos) q y p, situación que se da sólo con algunos pares. En este caso, q es un enlace entrante de p y éste uno saliente de q.

En particular, se ha estudiado la topología del grafo web [11] el cual se caracteriza por formar una red libre de escala el cual – además – es autosimilar, es decir, que porciones menores de éste mantienen propiedades del grafo completo [14]. Las redes libres de escala (scale-free network) se caracterizan por una distribución dispareja de nodos y enlaces [2]. Esto significa que se pueden encontrar nodos con muy pocos enlaces y otros con muchos. Los vínculos en páginas Web son un ejemplo de esto, tanto los entrantes como los salientes. En este caso, se observa en la Web que existen ciertos nodos que incorporan enlaces entrantes (crecen) de manera proporcional al tamaño que tienen. Estos nodos, resultan interesantes de encontrar y estudiar ya que vinculan partes importantes de la red. Kleinberg [16] y Barabasi [9] plantearon que la topología del grafo de la web corresponde a una red libre de escala, en la cual la distribución de los enlaces sigue una ley de potencias de la forma:

$P(x = k) \approx k^{-\beta}$, para $\beta > 0$, la cual expresa la probabilidad que la página x posea k enlaces.

El exponente β de la ley de potencias describe que tan rápido disminuye el valor de la frecuencia de x. Los ejemplos clásicos de estas distribuciones corresponden a Zipf y Pareto [1]. Esta situación fue luego observada por Broder en un muestreo de la web de gran escala [11], encontrando como propiedad básica del grafo web que la distribución del grado entrante de los vértices sigue una ley de potencias con exponente $\beta = 2.1$. Por otro lado, la distribución del grado saliente sigue una ley de potencias imperfecta con $\beta = 2.72$.

3 Metodología

Para el estudio de la web de Argentina se realizó una recolección de páginas utilizando el *crawler* WIRE [12] durante los meses de marzo y abril de 2006. Para delimitar el dominio de estudio se tomó como criterio recolectar sólo las páginas bajo el dominio '.ar'. Si bien se conoce que existen organizaciones de Argentina que utilizan el dominio '.com' para su sitio web, no es técnicamente simple obtener una lista exhaustiva de éstos y – además – varias de éstas mantienen el doble nombre de dominio, uno .com y otro. com.ar con redirección entre éstos en algunos casos o duplicación de contenidos en otros casos.

Bajo este criterio, el *crawler* fue inicialmente alimentado con aproximadamente 10.000 direcciones de dominios de Argentina obtenidos de directorios del país, de páginas oficiales gubernamentales y del directorio de Yahoo!. Para el análisis de los datos recolectados utilizamos la metodología propuesta en [8] estudiando la web en diferentes niveles de granularidad (páginas, sitios y dominios) y agrupando características de acuerdo a contenido, enlaces y tecnologías. No obstante, hemos agregado algunos estudios que permiten analizar características particulares del dominio en cuestión.

3.1 La colección WebAR

Se recolectaron 9.656.218 páginas desde 149.305 sitios que corresponden a 83.813 dominios. El 94.71% corresponden a páginas únicas y el 5.29% se encuentran duplicadas. Del total, el 48% corresponden a páginas estáticas, mientras que las dinámicas suman el 52%. Este es primer dato que nos llamó la atención y sobre el cual volveremos más adelante para proponer una explicación.

En febrero de 2006 solicitamos a NIC Argentina información acerca de la cantidad de dominios registrados bajo su administración. De acuerdo a la respuesta oficial [22], existen 1.129.381 dominios registrados y - según estudios propios acerca de utilización de los nombres de dominios - sólo el 26% (286.635) asignados bajo "com.ar" es productivo (publica un sitio web o recibe correo electrónico). En la tabla 1 se presenta la distribución de dominios de segundo nivel bajo ". ar", mientras que en la tabla 2 se indica la cantidad de dominios de tercer nivel, ajustados por el porcentaje de productividad en 'com.ar' y la cantidad de dominios pertenecientes a la muestra donde al menos se recuperó una página**.** Nótese que para los dominios diferentes de 'com.ar' se considera que el total de activos es cercano al 100% ya que las normas de registración son estrictas y se supone que no existen demasiados dominios sin utilizar.

Dominio de 2do nivel	Cantidad	%
com.ar	1.102.444	97,61
org.ar	14.133	1,25
net.ar	10.112	0,90
gov.ar	2.570	0,23
mil.ar	92	0,01
int.ar	30	0,00
Total	1.129.381	100

Tabla 1: Distribución de dominios de segundo nivel. Fuente: NIC Argentina, febrero 2006

Dominio 2do nivel	Dominios activos de 3er nivel (NIC)	Dominios activos de 3er nivel (Muestra)	% en la muestra
com.ar	286.635	77.668	27,10
org.ar	14.133	3.846	27,21
net.ar	10.112	817	8,08
gov.ar	2.570	896	34,86
mil.ar	92	21	22,83
int.ar	30	11	36,67
edu.ar (*)	N/D	554	
Total	313.572	83.813	

Tabla 2: Composición de la muestra. (*) El dominio .edu.ar no se encuentra bajo administración de NIC Argentina

4 Contenidos

Aquí se presentan los resultados del estudio del contenido en diferentes niveles de granularidad. A nivel de páginas se estudian las propiedades del texto mientras que para sitios y dominios se analizan cómo se distribuyen las páginas.

4.1 Tamaño de las Páginas

Por cada página descargada se almacenaron como máximo 100 Kb. Observamos que el tamaño medio de las páginas es de 10 Kb. Este valor es bastante menor que las observaciones de Chile (21 Kb) y Brasil (24 Kb). La distribución de los tamaños es muy sesgada y se puede modelar mediante una ley de potencias con parámetro $\beta = 2.2$ para las páginas cuyo tamaño es más de 20 Kb. En el gráfico 1 se presenta la distribución junto con la curva de ajuste.

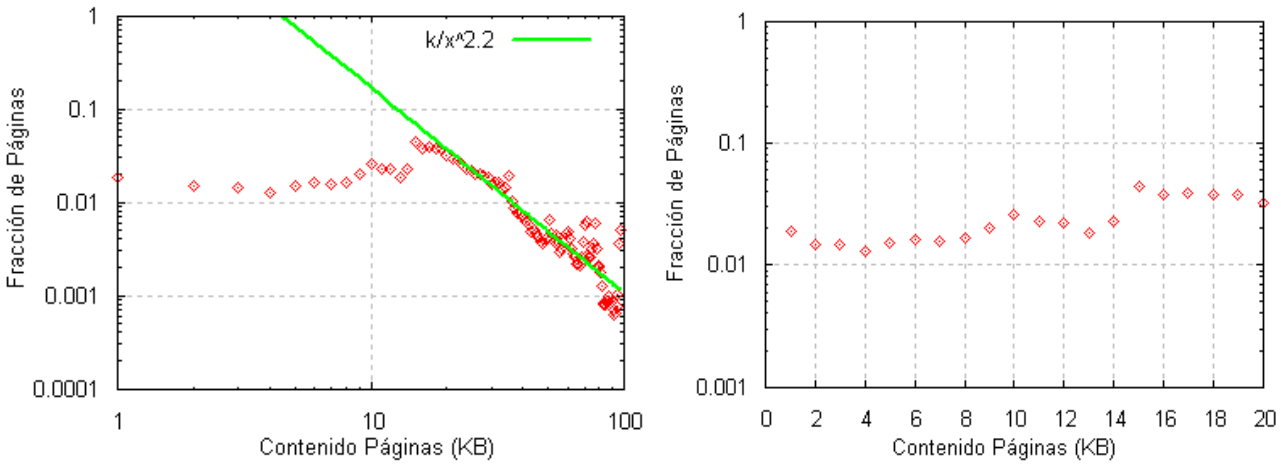

Gráfico 1 – Distribución de los tamaños de las páginas con la recta de ajuste (izquierda).
Detalle de la zona para los tamaños hasta 20 Kb (derecha)

4.2 Términos más Utilizados

Se extrajo de forma aleatoria un subconjunto de páginas para analizar su contenido. En total, se tomaron 396.134 documentos. Se eliminaron las marcas HTML y se extrajeron del texto puro todos sus términos. Luego, para cada documento se seleccionaron los 40 términos de mayor frecuencia. A continuación, se los fusionó en una única lista y se eliminaron palabras vacías en español e inglés y los términos de 1 carácter. Finalmente, se calculó el DF (*Document Frequency*), es decir, la cantidad de documentos en los que apareció cada uno, sin importar su TF (*Term Frecuency*) dentro de cada documento. En la tabla 3, se muestran los primeros 10 términos, ordenados por DF. En estos resultados se observa que los primeros lugares se encuentran ocupados por términos relacionados con la actividad comercial y – específicamente – con vocabulario propio de sitios dedicados a ventas masivas, subastas, catálogos en línea y demás.

4.3 Términos en los Nombres de Sitios

También realizamos un análisis de los nombres de sitios y dominios que conforman la muestra. No tomamos en cuenta las etiquetas de dominios de primer y segundo nivel, como tampoco interesó las palabra "www" por ser ampliamente utilizada. Por ejemplo, en el nombre de sitio "www.tyr.unlu.edu.ar" sólo analizamos la subcadena "tyr.unlu". Como separador de términos utilizamos el punto (".") y el guión medio ("-"). En la tabla 4 se muestran los primeros 10 términos más frecuentes en los nombres de sitios. A diferencia del análisis de contenido, en los nombres de sitios comienzan a aparecer términos relacionados con el turismo (itálica). Una observación interesante es que algunos de estos sitios poseen su nombre de dominio formado por un término concatenado con la palabra “argentina”. Por ejemplo, hoteleinrgentina, hotelesargentina, viajeaargentina, viajarxargentina y alojarseargentina, entre otros. Entre las primeras 100 palabras más utilizadas, el nombre Argentina aparece 12468 veces (4.6%).

Orden	Término	Cantidad de Documentos	%
1	precio	67,966	17.16
2	compra	67,456	17.03
3	inicio	60,362	15.24
4	artículos	59,831	15.10
5	venta	58,930	14.88
6	argentina	56,944	14.37
7	cuotas	50,047	12.63
8	tarjeta	49,926	12.60
9	comprar	46,824	11.82
10	pagofacil	46,729	11.80

Tabla 3 – Primeros 10 términos más utilizados en el contenido de los documentos

Orden	Término	Cantidad de Sitios
1	*campings*	51,318
2	*sbviajes*	21,922
3	*argentina*	7,384
4	*tango*	7,256
5	*europa*	6,835
6	*brasil*	6,472
7	*aereos*	6,215
8	*paquetes*	6,194
9	noticias	6,174
10	ofertas	6,103

Tabla 4 – Primeros 10 términos más frecuentes en los nombres de sitios

4.4 Páginas por Sitio

La cantidad media de páginas por sitio es 65 y su distribución corresponde a una ley de potencias con parámetro β = 1.45 (Gráfico 2). Este valor es comparable con otros países de la región como Chile donde se encontraron 58 [5] con una distribución de parámetro β = 1.6 o Brasil con 66 [18] y β = 1.6. Inclusive, resulta similar a la web de España que posee una media de 52 páginas y β = 1.1 [6].

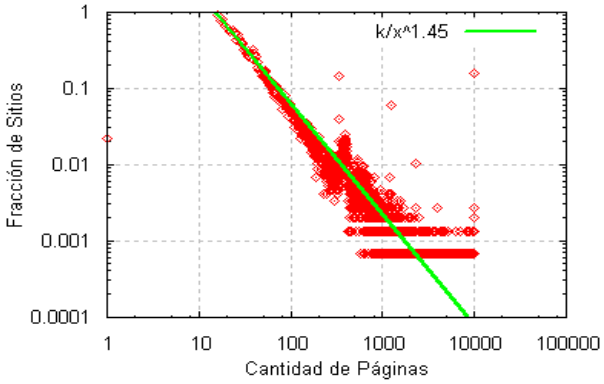

Gráfico 2 – Distribución de la cantidad de páginas por sitio.

5 – Enlaces

Esta sección corresponde al estudio de las relaciones establecidas a nivel de enlaces. Como se mencionó anteriormente, el espacio web es modelado como un grafo dirigido sobre el cual se analizan características.

5.1 Grado Entrante y Saliente de las Páginas

El grado entrante de una página corresponde al número de enlaces desde otras que apuntan hacia ésta, mientras que el grado saliente corresponde al número de enlaces que posee una página hacia otras. La distribución de grado entrante es bastante sesgada y sigue una ley de potencias con parámetro β = 1.71 (Gráfico 3). En Brasil, se encontró β = 1.0, en Chile β = 2.0 y en España β = 2.1. Como dato interesante, hallamos un 55% de páginas con grado entrante igual a cero.

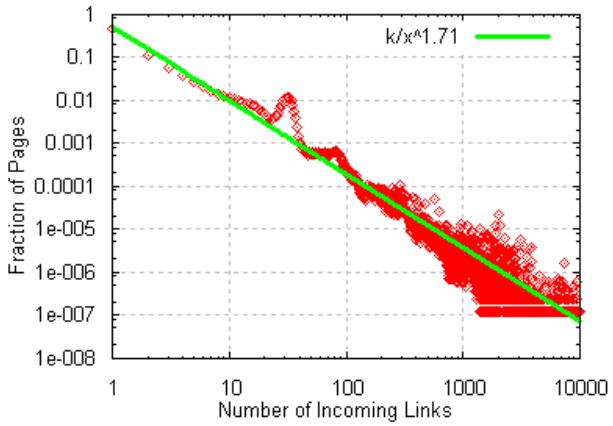

Gráfico 3 – Distribución del grado entrante de las páginas

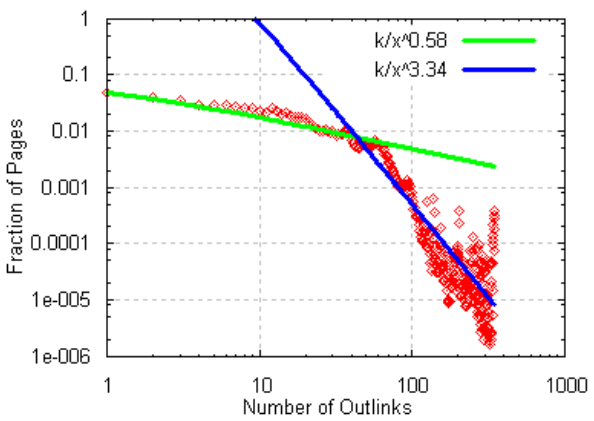

Gráfico 4 – Distribución del grado saliente de las páginas con las dos curvas de ajuste

Por otro lado, la distribución de grado saliente es mas sesgada que la anterior y se pueden analizar usando dos leyes de potencias [8], lo que permite realizar una mejor aproximación. La primera para valores menores que un umbral y la segunda para los superiores (Gráfico 4). En este caso, obtuvimos un ajuste para valores menores a 30 enlaces salientes con exponente $\beta_1 = 0.58$ y para la correspondiente a los valores superiores el ajuste fue con $\beta_2 = 3.34$. Estas distribuciones son comparables con las halladas para Chile ($\beta_1 = 0.7$ y $\beta_2 = 2.6$), Brasil ($\beta_1 = 0.7$ y $\beta_2 = 2.7$) y España ($\beta_1 = 0.9$ y $\beta_2 = 4.2$). Sobre esta característica hallamos que aproximadamente el 30% de las páginas no poseen enlaces salientes.

5.2 PageRank

El ranking de las páginas a partir del análisis de enlaces es una característica importante a estudiar ya que algunos motores de consultas utilizan información del grado de las páginas para establecer la importancia de cada una de éstas. Esta idea se fundamenta en que la estructura de enlaces es armada – en general – por humanos y representa una fuente de información indirecta (respecto del contenido) que es de alto valor [13]. Esta información es utilizada en diversas aplicaciones como búsquedas, minería web y ranking (por ejemplo, mediante HITS [17] y PageRank [20]). En particular, calculamos los valores de Pagerank por ser uno de los algoritmos más citados. En el gráfico 5 se aprecia que la distribución sigue una ley de potencias con parámetro β = 1.85. En otros estudios se hallaron distribuciones similares, como en Chile (β = 1.9), Brasil (β = 1.8) y España (β = 2.0). De acuerdo a [21] este exponente debería ser similar al de la distribución de grado entrante, siyuación que se da en este caso.

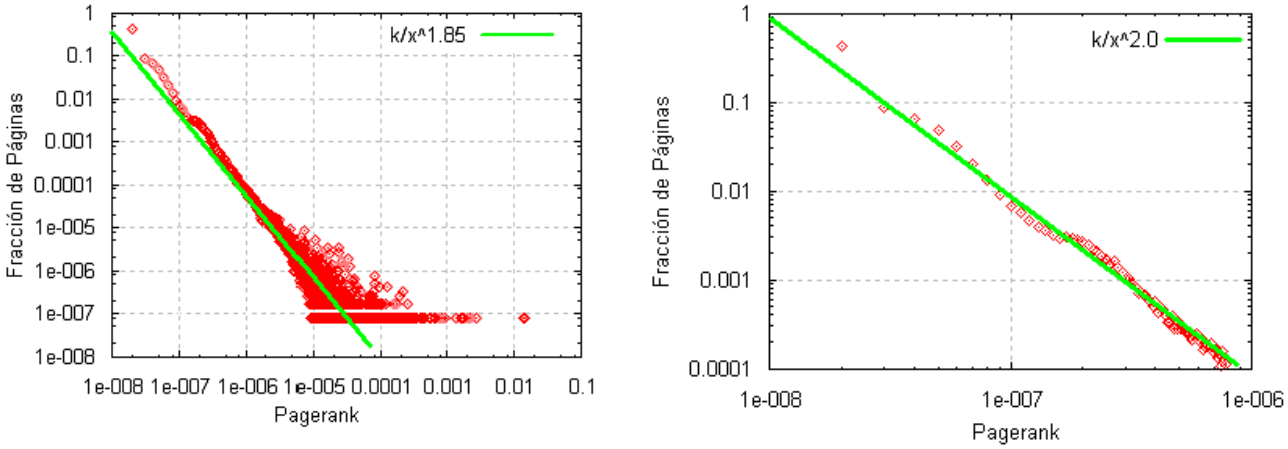

Gráfico 5 – Distribución de las páginas por valor de Pagerank (izquierda). Ajuste para los primeros valores de Pagerank que agrupan la mayor cantidad de las páginas (derecha) con β = 2.0.

5.3 Grado Entrante y Saliente en Hostgraph

Se denomina *hostgraph* al grafo creado cambiando los nodos que representan páginas web en el mismo sitio por uno único que representa el sitio web completo [8]. Luego, si existe al menos un enlace de una página de un sitio a otra en otro, entonces existirá un enlace a nivel de Hostgraph. En este caso, la distribución de enlaces entrantes se ajusta a una ley de potencias con parámetro β = 1.7 (Gráfico 6). Situaciones similares se encontraron en Chile (β = 2.0), Brasil (β = 1.9) y España (β = 1.8). De manera complementaria, se calculó la distribución de enlaces salientes en Hostgraph la cual también corresponde a una ley de potencias con parámetro β = 1.5 (Gráfico 7). Comparando con otros países encontramos: Chile (β = 1.9), Brasil (β = 1.9) y España (β = 1.3).

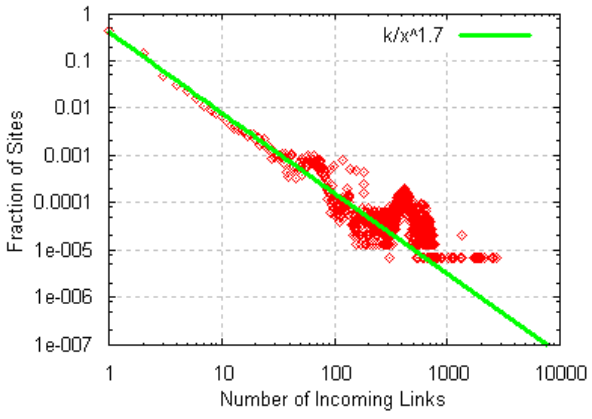

Gráfico 6 – Distribución de grado entrante en *Hostgraph*

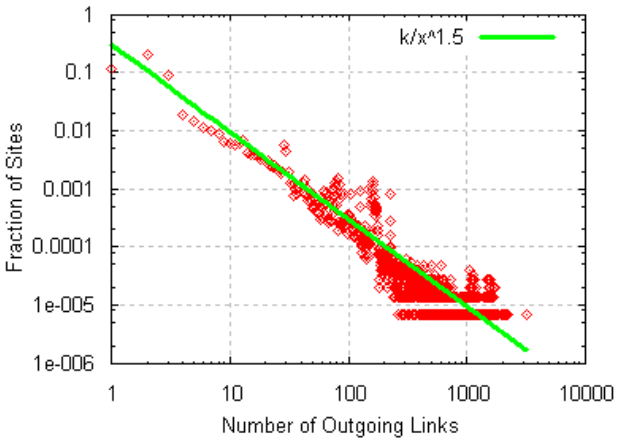

Gráfico 7 – Distribución de grado saliente en *Hostgraph*

5.4 Componentes Fuertemente Conectados

Se estudió la distribución de los Componentes Fuertemente Conectados (SCC) del grafo a nivel de sitios. Un SCC es un subgrafo dirigido en el cual todos los nodos pueden alcanzar a los demás (dentro del mismo subgrafo) siguiendo los enlaces. En la muestra de la web argentina podemos observar la existencia de una componente gigante (Tabla 5), mientras que la distribución de los tamaños sigue una ley de potencias con $\beta = 2.74$ (Gráfico 8).

Tamaño	Número	Tamaño	Número
1	66.021	12	2
2	432	14	1
3	81	16	2
4	164	20	1
5	18	21	1
6	9	22	1
7	8	23	1
8	1	29	1
9	4	38	1
10	2	44	1
11	2	80.968	1

Tabla 5 – Tamaño de los SCCs

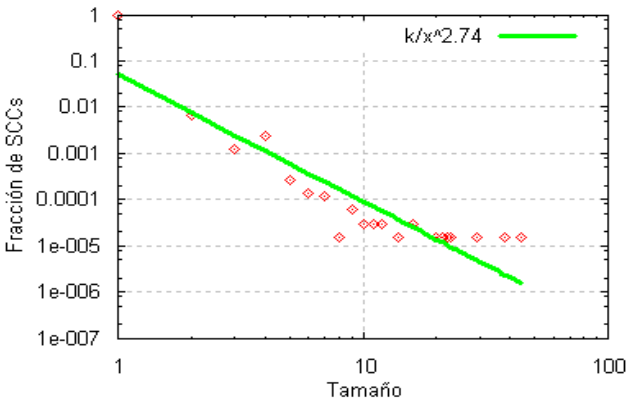

Gráfico 8 – Distribución de los Tamaños de los SCCs

5.6 Estructura Macroscópica

En [11] se propuso una estructura que muestra las relaciones existentes entre las páginas respecto de un subconjunto de grafo web correspondiente al SCC de mayor tamaño. Esta estructura, conocida como "*bow-tie*", ubica a cada página en una de 6 regiones: a) **MAIN**, que incluye al SCC de mayor tamaño; b) **IN**, formado por nodos que pueden alcanzar a los nodos en MAIN pero no son alcanzables desde éste; c) **OUT**, que es un conjunto de nodos alcanzables desde MAIN que no poseen enlaces salientes hacia éste; d) **ISLANDS**, nodos desconectados de los los componentes anteriormente mencionados; e) **TENTACLES**, nodos que son alcanzables solamente desde porciones de IN o de OUT y f) **TUNNELS**, nodos desde IN que alcanzan a otros en OUT. En [3] se extendió en nivel de detalle del componente MAIN dividiéndolo en las siguientes subregiones: a) **MAIN-MAIN**, sitios que pueden ser alcanzados directamente desde la componente IN o que pueden alcanzar directamente la componente OUT; b) **MAIN-IN**, sitios que pueden ser alcanzados directamente desde IN pero no están en MAIN-MAIN; c) **MAIN-OUT**, sitios que pueden alcanzar directamente a OUT y no pertenecen a MAIN-MAIN y d) **MAIN-NORM**, sitios que no pertenecen a las anteriores.

El tamaño de la región MAIN (54,23%) muestra que la web argentina se encuentra – en general – bien conectada, especialmente si lo comparamos con Chile y Brasil donde el porcentaje de sitios es 21.76 y 25.27%, respectivamente. Los sitios en la componente OUT (28.15%) representan una fracción más baja que en Brasil (45.33%) pero similar a Chile (26.12%). Si se tiene en cuenta que uno de los motivos que hacen que un nodo esté en OUT es su antigüedad y desactualización, podemos ver que en ".ar" no representan un porcentaje mayor. Por otro lado, los sitios pertenecientes a las componentes IN e ISLANDS únicamente se los accede a partir de sus páginas iniciales debido a que pueden ser páginas nuevas o no estar bien conectadas. En este caso, representan porcentajes comparables con Chile en la componente IN (6.65%) pero no en ISLANDS (46.16%). En Brasil se reportaron un 12.95% y 12.35% respectivamente. En la tabla 6 se presentan los resultados completos.

Componente	Sitios	%
MAIN	80.968	54,23
MAIN_NORM	*50.346*	*33,72*
MAIN_MAIN	*10.212*	*6,84*
MAIN_IN	*3.439*	*2,30*
MAIN_OUT	*16.971*	*11,37*

Componente	Sitios	%
IN	8.523	5,71
OUT	42.026	28,15
TIN	2.915	1,95
TOUT	951	0,64
TUNNEL	176	0,12
ISLAND	13.746	9,20

Tabla 6 – Componentes de la estructura macroscópica

6 Tecnologías

En esta sección se presentan los resultados de estudios relacionados con las tecnologías utilizadas para la gestión de la información publicada, como distribuciones de formatos de archivos y lenguajes de programación, entre otros.

6.1 Códigos de Respuestas HTTP

En primer lugar se muestran los resultados de las respuestas entregadas por los servidores durante la etapa de recolección de páginas. De manera normal, un cliente web (en este caso el *crawler* WIRE) abre una conexión TCP con el servidor web correspondiente y solicita – mediante el protocolo HTTP – el recurso deseado. El servidor responde con un código de estado. La evaluación de los mismos permite determinar si la página se puede descargar y los diferentes motivos por los cuales no se puede recuperar. Los códigos HTTP se agruparon en:

- **OK:** Incluye todos los pedidos exitosos: OK (200) y PARTIAL CONTENT (206).
- **MOVED:** Agrupa los pedidos en los cuales el servidor informa acerca de una redirección de la página solicitada: MOVED (301), FOUND (302) y TEMPORARY REDIRECT (307).
- **SERVER ERROR:** Coresponde a todas las peticiones fallidas en el lado del servidor: INTERNAL SERVER ERROR (500), BAD GATEWAY (502), UNAVAILABLE (503), y NO CONTENT (204).
- **FORBIDDEN:** Agrupa todas las peticiones que no son permitidas por el servidor: UNAUTHORIZED (401), FORBIDDEN (403) y NOT ACCEPTABLE (406).
- **NOT FOUND:** Representan el código de mismo nombre (404).
- **OTHER.**

En la tabla 7 y en el gráfico 9 se presentan los resultados sobre la base de 12.276.090 páginas solicitadas. Como se puede apreciar, el porcentaje de descargas exitosas es bueno y se encuentra dentro de los valores reportados en otros estudios que están entre el 75 y 85% [8]. Sin embargo, la proporción de enlaces rotos (más del 5%), es significativa. Esto indica algún problema relacionado con el mantenimiento de los documentos de un dominio lo cual, teniendo en cuenta la disponibilidad de herramientas para el chequeo automático, se podría minimizar o eliminar.

Grupo	Cantidad
OK	9.656.218
MOVED	952.196
OTHER	897.436
NOT FOUND	655.114
SERVER ERROR	61.613
FORBIDDEN	53.513
TOTAL	12.276.090

Tabla 7 – Distribución de los Códigos HTTP

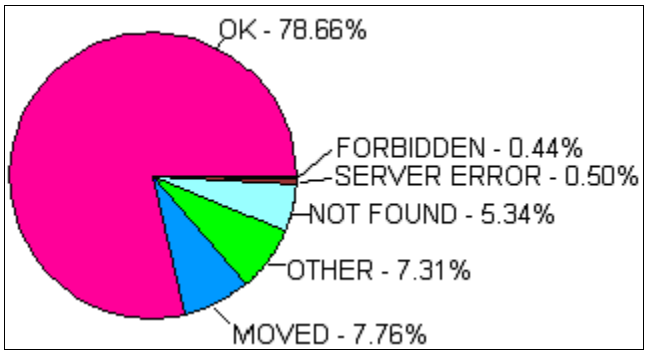

Gráfico 9 – Distribución de los Códigos HTTP

6.2 Longitud de las URLs

Se estudió la distribución de la longitud (en bytes) de las URLs, la cual se presenta en el gráfico 10. Observamos una longitud promedio de 68 bytes sin incluir la parte correspondiente al protocolo, lo que la incrementaría en 7 bytes (http://). Este valor es comparable similar a los observados en Chile (64 bytes), Brasil (69 bytes) y España (67 bytes). Encontramos – además – que en 160 bytes se encuentra el 99% de las URLs y en 100 bytes el 92%. Por otro lado, observamos URLs muy largas (hasta 1000 bytes) las cuales cuentan un porcentaje menor (menos del 0.5%) y corresponden a páginas dinámicas.

Complementariamente, estudiamos la longitud de las URLs de páginas HTML (estáticas) y dinámicas y los parámetros de éstas. En el gráfico 11 se presentan las distribuciones correspondientes. En el primer caso, el ajuste corresponde a una distribución normal con parámetros $\mu = 45$ y $\sigma = 10$, mientras que para las páginas dinámicas el ajuste también es normal con $\mu = 60$ y $\sigma = 19$. Aquí se puede observar que con 110 bytes se obtienen el 99% de las las páginas HTML y el 99.8 de las páginas dinámicas. El promedio de longitud para las páginas HTML que observamos es de 62 bytes,

mientras que para las páginas dinámicas es de 46 bytes. Esta diferencia puede atribuirse a que en los sitios manejados de forma dinámica no utilizan demasiadas jerarquías de directorios. Por otro lado, en las páginas dinámicas, la longitud promedio de los parámetros es de 27 bytes.

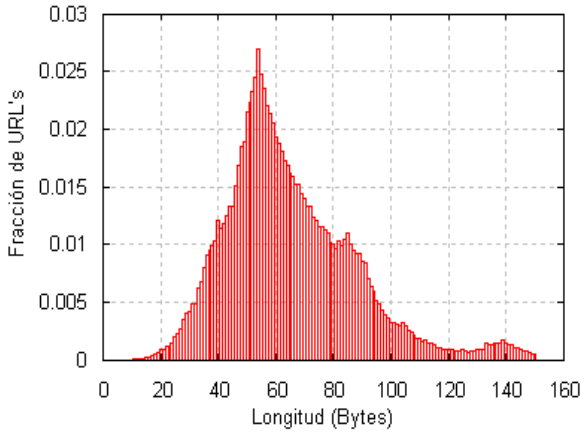

Gráfico 10 – Distribución de las longitudes de las URLs.

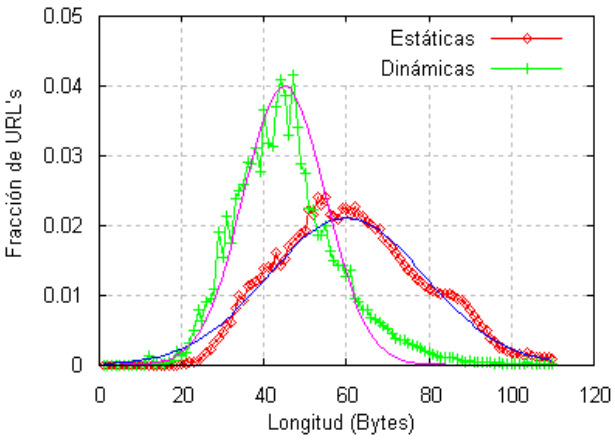

Gráfico 11 – Distribución de las longitudes de las URLs de páginas estáticas con ajuste N(45, 10) y dinámicas con ajuste N(60, 19)

6.3 Profundidad de los Documentos

La profundidad de un documento es el número de enlaces que es necesario seguir desde el inicio de un sitio para alcanzarla. El inicio o portada de un sitio posee profundidad 0, las páginas directamente alcanzables desde el inicio profundidad 1, y así sucesivamente. Se limitó al módulo recolector para que descargue solamente 5 niveles para páginas dinámicas, y sólo 15 niveles para páginas estáticas. El máximo se sitúa en el nivel cuatro (tabla 8 y gráfico 12).

Profundidad	Documentos	%
1	532.003	4,33
2	1.350.455	11,00
3	4.154.504	33,84
4	4.964.279	40,44
5	1.161.936	9,47
6	86.238	0,70
7	22.189	0,18
8	3.679	0,03
9	804	0,01
10	3	0,00
Total	12.276.090	100,00

Tabla 8 – Distribución de los documentos por profundidad

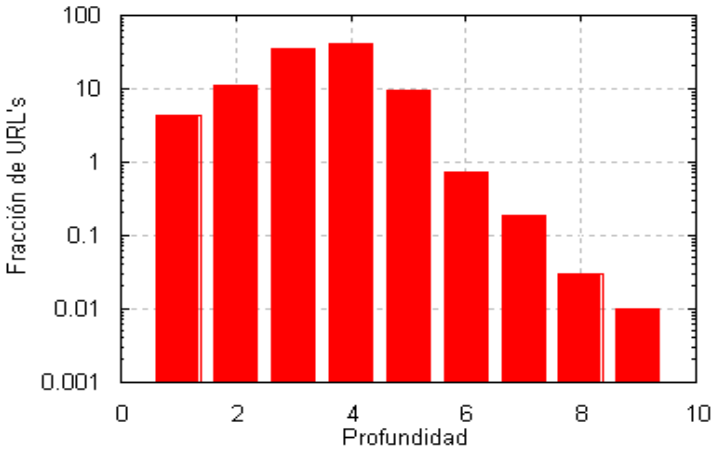

Gráfico 12 – Distribución de la proporción de documentos por profundidad (eje y en escala logarítmica)

6.4 Documentos Estáticos vs Dinámicos

Para este estudio dividimos los documentos descargados en dos grupos, tratando de identificar aquellos que se denominan "páginas dinámicas". Éstas son páginas HTML que no se encuentran previamente almacenadas en el disco rígido del servidor web sino que son confeccionadas "*on the fly*" por un programa y entregadas al cliente. Generalmente, las páginas dinámicas se arman para entregar respuestas a consultas a bases de datos, a determinadas opciones ingresadas por los usuarios ó para armar sitios personalizados.

En el análisis de documentos estáticos y dinámicos observamos una llamativa paridad (Tabla 9), inclusive con una superioridad de páginas dinámicas (52%). Si comparamos con España y Chile, cuyas muestras presentan un 22% y 38%

de páginas dinámicas respectivamente, este porcentaje es bastante elevado. Estos valores denotan que existe una importante infraestructura de desarrollo web que soporta gran parte de la lógica de negocios de las organizaciones. Otra posible explicación podría deberse a que tanto la web de Chile como la de Argentina se suponen más nuevas que la de España, por lo que la utilización de tecnologías dinámicas tiene una mayor desarrollo en los últimos años.

Complementariamente, analizamos la distribución de los enlaces a documentos con las extensiones utilizadas para construir páginas dinámicas (Gráfico 13). Se puede apreciar una importante participación del lenguaje de preprocesamiento de hipertextos PHP con un 52% seguido por Perl con 39%, ambas herramientas completamente libres de costo de utilización. En cuanto a España, hay aproximadamente un 46% de uso de PHP, pero lo sigue un 44% de ASP, mientras que en Chile hay un 78% de PHP y un 16% de ASP. Por otro lado, Brasil cuenta con más del 70% y 20% respectivamente. En estos 3 países la utilización de Perl es proporcionalmente muy baja.

	Documentos	%
TOTAL	12.276.090	100,00
Dinámicos	6.383.050	52,00
Eestáticos	5.893.040	48,00

Tabla 9 – Distribución de documentos estáticos y dinámicos

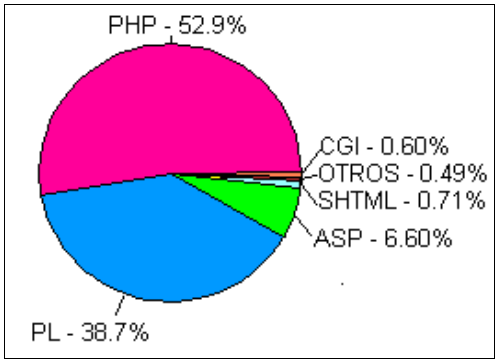

Gráfico 13 – Distribución de enlaces a documentos con extensiones de páginas dinámicas

6.5 Distribución de Sitios por País

Sobre una muestra extraída al azar, compuesta por 23.965 nombres de sitios (que representa el 16.05% del total de sitios donde el módulo de *crawling* recuperó al menos una página web) se evaluó en qué países se hallan hospedados los sitios que almacenan contenidos del dominio ".ar. A los efectos de relacionar direcciones de red con países se utilizó la bases de datos geográficos GeoIPCountryWhois de la empresa Maxmind[1].

Países	Sitios	%	Países	Sitios	%
Argentina	18.177	75,87	**Reino Unido**	60	0,25
Estados Unidos	4.700	19,62	**Israel**	48	0,20
Canadá	351	1,47	**Lituania**	39	0,16
Brasil	224	0,94	**Chile**	6	0,03
Colombia	150	0,63	**Alemania**	5	0,02
España	89	0,37	**Otros paises**	24	0,10
Francia	84	0,35	**TOTAL**	23.957	100,00

Tabla 10 – Distribución de sitios por país

A partir de estos datos (Tabla 10) se puede observar que casi el 76% de los sitios se hallan alojados en servidores que residen en la República Argentina. Entendemos que este dato es un indicador más acerca del desarrollo tecnológico del país en estudio. Por otro lado, para los usuarios argentinos que deseen contratar servicios de alojamiento de sitios en el exterior existe una diferencia económica significativa debido a la paridad entre la moneda nacional y el dólar o el euro.

7 Conclusiones

En este trabajo se presenta una caracterización del espacio web de Argentina sobre una muestra propia de 9.656.218 páginas pertenecientes a 149.305 sitios en 83.813 dominios de tercer nivel. Para el análisis, se dividió el estudio en

[1] http://www.maxmind.com/

cuenta a contenido, enlaces y tecnologías utilizadas. En primer lugar, observamos una alta participación de sitios bajo el dominio ".com.ar", inclusive sólo considerando aquellos que se encuentran activos (26%), de acuerdo a información de NIC Argentina, organismo oficial de registración. Una posible cuestión a tener en cuenta es el estudio de replanteo de los mecanismos de asignación y registro de nombres de dominio. Según los datos obtenidos gran parte de los dominios "com.ar" no están siendo utilizados.

En cuanto al contenido, se observó que la distribución de los tamaños de las páginas es bastante segada. En el estudio del vocabulario de las páginas se encontró que predominan términos relacionados con la actividad comercial como sitios dedicados a ventas masivas, subastas, catálogos en línea y demás. Esta situación puede acarrear problemas de pérdida de precisión en ciertos tipos de búsquedas debido a que – generalmente – estos sitios están muy bien posicionados en los rankings. Sin embargo, en los nombres de los sitios, extraídos de las URLs, aparecen mayormente términos relacionados con el turismo, actividad de mucho auge en los últimos años en Argentina.

Del análisis de enlaces y conectividad surge muestra que la web argentina se encuentra – en general – bien conectada. Un indicador es que la componente MAIN posee el 54.23% de los sitios, mientras que hay una baja proporción en ISLANDS (9,21%), lo que refuerza esta idea. Esta situación se mantiene inclusive al analizar los sitios por dominio de segundo nivel. Por otro lado – y como se esperaba – las distribuciones enlaces entrantes, salientes y pagerank siguen leyes de potencias. Una alta proporción de las páginas (55%) no posee enlaces provenientes de otros sitios del dominio ".ar" y un 30% no poseen enlaces salientes.

En cuanto a los aspectos tecnológicos, hallamos que – del total de páginas descargadas – el 48% son estáticas y el 52%, dinámicas, las cuales se encuentran construidas en gran parte utilizando herramientas libres como PHP (53%) y Perl (39%). Además, casi el 76% de los sitios se hallan alojados en servidores que residen en Argentina y el 68% de las direcciones de red donde se alojan sitios web están en el país. De estos indicadores se desprende que existe un importante desarrollo tecnológico y de la infraestructura de comunicaciones de Argentina relacionada con la web.

De este estudio de desprenden varias líneas de investigación y desarrollo. En primer lugar, consideramos interesante realizar nuevos trabajos que permitan armar mapas de evolución y dinámica del espacio objeto a los efectos de estudiar su comportamiento en el tiempo. Además, surge la necesidad de construir servicios de información locales que utilicen la información obtenida para mejorar la experiencia de los usuarios con aplicaciones basadas en el contenido de la web, por ejemplo, permitiendo filtrar las respuestas provenientes de sitios comerciales. Por otro lado, el estudio en profundidad de porciones más acotadas como – por ejemplo – el dominio edu.ar permitiría obtener indicadores de desarrollo para la comunidad educativa.

Agradecimientos

Agradecemos al Ingeniero Jorge Vilas de NIC Argentina por la información aportada en cuanto a la cantidad de dominios registrados. También al personal técnico de RETINA por su valiosa colaboración.

Referencias

[1] L.A. Adamic and B.A. Huberman. Zipf's law and the Internet. Glottometrics 3, pp. 143-150, 2002.

[2] R. Albert R. and A.-L. Barabasi. Statistical mechanics of complex networks. Review of Modern Physics 74, pp. 47-94, 2002.

[3] R. Baeza-Yates and C. Castillo. Relating Web characteristics with link based Web page ranking. In Proceedings of String Processing and Information Retrieval (SPIRE), IEEE Cs. Press, pp. 21-32. Laguna San Rafael, Chile, 2001.

[4] R. Baeza-Yates and F. Lalanne. Characteristics of the Korean Web. Technical Report, Korea-Chile IT Cooperation Center, ITCC, 2004.

[5] R. Baeza-Yates and C. Castillo. Características de la Web Chilena 2004. Technical Report, Center for Web Research, University of Chile, 2005.

[6] R. Baeza-Yates, C. Castillo and V. Lopez. Characteristics of the Web of Spain. Cybermetrics, Vol. 9, Nro. 1, 2005.

[7] R. Baeza-Yates, and C. Castillo. Link Analysis in National Web Domains. Workshop on Open Source Web Information Retrieval (OSWIR), pp. 15-18. Compiegne, France, 2005.

[8] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of national Web domains. Technical report, Universitat Pompeu Fabra, July 2005.

[9] A. L. Barabasi and A. Albert. Emergence of Scaling in Random Networks. Science, (286): pp. 509-512, 1999.

[8] K.Bharat, B-W. Chang, M. Herzinger and M. Rhul. Who Links to Whom: Mining Linkage between Web Sites. In Proceedings of the IEEE International Conference on Data Mining, 2001.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph Structure in the Web. In Proceedings of the WWW9 Conference pp. 309-320, 2000.

[12] C. Castillo and R. Baeza-Yates. WIRE: an Open Source Web Information Retrieval Environment. Workshop on Open Source Web Information Retrieval (OSWIR), 2005.

[13] S. Chakrabarti, B.E. Dom, D. Gibson, D., and J. Kleinberg. Mining the Link Structure of the World Wide Web. IEEE Computer, Vol. 32, No. 8, pp: 60-67, 1999.

[14] S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. ACM Transactions on Internet Technology, Vol. 2, Nro.3, pp. 205-223, 2002.

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

[16] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a Graph: Measurements, Models and Methods. In Proceedings of the International Conference on Combinatorics and Computing, 1999.

[17] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Association for Computing Machinery - Journal of the Association for Computing Machinery, Vol. 46, Nro. 5, pp. 604-632, 1999.

[18] M. Modesto, A. Pereira, N. Ziviani, C. Castillo and R. Baeza-Yates. Un Novo Retrato da Werb Brasileira. In Proceedings of SEMISH, São Leopoldo, Brazil, 2005.

[19] E. O'Neill, B. Lavoie, R. Bennett. Trends in the Evolution of the Public Web 1998 - 2002. D-Lib Magazine, Vol. 9, Nro. 4, 2003.

[20] L. Page, S. Brin, R. Montwani and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library Technologies Project, 1998

[21] G. Pandurangan, P. Raghavan, and E. Upfal. Using Pagerank to characterize Web structure. In Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON), volume 2387 of Lecture Notes in Computer Science, pp. 330-390, Singapore, 2002.

[22] J. Vilas. Solicitud de datos para investigación sobre "Caracterización de la web Argentina". Comunicación Personal, Febrero 6, 2006.

[23] Web Characterization Activity. http://www.w3.org/WCA/