El Estado de la Web de Paraguay y la Sociedad de la Información

Fernando R. A. Bordignon
Universidad Nacional de Luján.
Departamento de Ciencias Básicas
Laboratorio de Redes
bordi@unlu.edu.ar

Pablo J. Lavallén
Universidad Nacional de Luján.
Departamento de Ciencias Básicas
Laboratorio de Redes
plavallen@unlu.edu.ar

Gabriel H. Tolosa
Universidad Nacional de Luján.
Departamento de Ciencias Básicas
Laboratorio de Redes
tolosoft@unlu.edu.ar

Resumen

Este trabajo tiene por objetivo principal la caracterización del espacio web de Paraguay, en el marco de las nuevas tendencias de crecimiento y evolución. Dado que es un sistema basado en interacciones sociales una mejor compresión de su naturaleza permite establecer patrones y tendencias para diseñar mejores estrategias que faciliten el acceso.

Como datos más significativos aportados por el presente estudio se citan los siguientes: a) El espacio web de Paraguay posee más de 2.800 sitios que corresponden a más de 2.600 dominios activos. En tales sitios se recolectaron más de 210.000 páginas. Existe cerca de un 5,5% de enlaces que no existen. b) El promedio de páginas por sitio es de alrededor de 79 documentos y el tamaño promedio del sitio es de 1,49 MB. c) Según el análisis de macroestructura de Broder se observa que el 55% de los sitios son islas es decir no intercambian enlaces con el resto de las componentes. d) El 60% de las páginas de la web Paraguaya fue creada o actualizada en el último año, lo cual indica una importante dinámica. e) El 93,8% de los enlaces externos son a dominios com (70,01%) y py (23,7%) y f) El 63,14% de las páginas recolectadas corresponden a páginas estáticas, mientras que las dinámicas suman el 36,86%.

1 – Introducción

El concepto de Sociedad de la Información (SI) se ha utilizando en un amplio espectro de contextos, desde la visión del desarrollo tecnológico y la masificación de los servicios brindados por las redes globales, hasta los planes de los gobiernos para la incorporación de los conceptos relacionados con el acceso masivo a la información como otra una base del desarrollo de una sociedad. Si bien se pueden encontrar definiciones más o menos exhaustivas, en general, se acepta la idea de “Sociedad del la Información” como un elemento más en el análisis del desarrollo social general de un país o región. Ésta, se encuentra caracterizada por la capacidad de sus miembros – desde individuos hasta organizaciones y gobiernos – para acceder y compartir información desde cualquier lugar, por medios electrónicos. Además, este grado de desarrollo genera oportunidades diferentes entre los individuos, lo que se conoce como “brecha digital”. En un modelo propuesto en [Redondo, 2004] se distinguen cuatro actores principales en la SI, a saber:

a) **Entorno**: factores o agentes que influyen en la sociedad como – por ejemplo - legislación, formación y cultura, los cuales indefectiblemente también influyen en la SI.
b) **Infraestructuras**: Recursos tecnológicos que brindan acceso las comunicaciones y – por ende – a compartir y utilizar contenidos.
c) **Contenidos**: son los recursos lógicos disponibles a través de los medios tecnológicos, desde información hasta servicios en línea, los cuales son utilizados o accedidos desde cualquier parte.
d) **Usuarios**: Son los individuos y las organizaciones que utilizan los contenidos mediante el acceso a través de las infraestructuras disponibles.

Aquí se nota la estrecha relación entre las componentes, lo cual genera una dependencia directa. Además, los dos primeros influyen sobre el desarrollo de los contenidos y sobre las posibilidades de los usuarios. Esto significa que existe una necesidad real de balance de recursos, con objetivos claros y concretos.

Es reconocible la participación y la importancia de Internet en este proceso, no solamente por su alcance global, sino – además – por la posibilidad que brinda de conocer e interactuar entre sociedades distantes, tanto geográfica como culturalmente. Complementariamente, el estadío de desarrollo de Internet en una región es parte del Índice de la Sociedad de la Información (ISI), indicador propuesto por la International Data Corporation[1] [Menezes, 2001]. De hecho, la UNESCO ha planteado la evolución hacia las Sociedades del Conocimiento a partir del auge de Internet y demás tecnologías digitales [UNESCO, 2005].

En este contexto, la World Wide Web tiene un rol preponderante. Desde que se ha convertido en el servicio de mayor expansión en la última década presenta características cada vez más completas y complejas. Hoy en día, la web es parte de las bibliotecas (o una biblioteca más) y – por ende - es patrimonio de las naciones.

En la actualidad el ciudadano es parte de una revolución basada en el uso y manejo de la información disponible principalmente en la web, por lo que es necesario conocer su grado de desarrollo, su extensión y calidad. De acuerdo a Baeza-Yates [Baeza-Yates, 2003], "*La Web no sólo crece, también evoluciona y además parte de ella muere*". Si bien, originalmente, la web 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. De acuerdo a David Sifry, fundador de Technorati[2], la Internet actual es más social y más participativa a usuarios que publican [EL PAIS, 2006]. Además, las últimas tendencias hacia la Web Semántica y Web 2.0 ya están marcando el rumbo de los próximos años.

La caracterización de espacios web entrega una valiosa información, tanto para entender su estructura como para desarrollar herramientas que faciliten la utilización de sus recursos. Esta es una tarea compleja, que requiere de recursos de gran escala debido a su tamaño y distribución geográfica. Se han realizado estudios sobre dominios nacionales

[1] http://www.idc.com/
[2] http://www.technorati.com/

[Baeza-Yates, 2004] [Baeza-Yates, 2005_a] [Baeza-Yates, 2005_b] [Efthimiadis, 2004] [Modesto, 2005], tomando muestras de dominios variados y utilizando diferentes estrategias de recolección. De acuerdo a [Baeza-Yates, 2005_c], estas muestras presentan un buen balance entre diversidad y completitud, por lo que constituyen un conjunto de alto interés.

Este trabajo tiene por objetivo principal la caracterización del espacio web de Paraguay, en el marco de las nuevas tendencias de crecimiento y evolución. Dado que es un sistema basado en interacciones sociales una mejor compresión de su naturaleza permite establecer patrones y tendencias para diseñar mejores estrategias que faciliten el acceso. Desde el punto de vista del modelo de la SI presentado, se estudian los Contenidos y su estructura lógica, lo que permitirá obtener indicadores que posibiliten la planificación de acciones del Entorno con el objetivo final de expandir el acceso y la información a los Usuarios.

Complementariamente, se desarrolló un modelo de visualización de este espacio web basado en el concepto de *tag-clouds* [Godwin-Jones, 2006], extendido a sitios web, lo que denominamos *site-clouds* (nube de sitios). En el mismo, se presentan – en una dimensión – los sitios más importantes de acuerdo a diferentes criterios en una nube y luego, los términos más representativos que lo definen. De esta manera, se pretende mostrar visiones distintas acerca de los sitios que componen el espacio en estudio. La aplicación se encuentra disponible en http://www.tyr.unlu.edu.ar/research/webpy/.

Concretamente, la principal contribución de este trabajo es:

a) Una caracterización del espacio web de Paraguay a partir de una muestra obtenida en julio de 2006 (altamente actualizada).
b) Una aplicación de visualización alternativa de la misma basada en *tag-clouds* pero extendida al nivel de sitios (*site-clouds*).
c) Sugerencias acerca de las características principales a desarrollar en este espacio de estudio a los efectos de aumentar los recursos ofrecidos a la comunidad.

El artículo se encuentra organizado de la siguiente manera. En la sección 2 se explica el proceso de recolección de datos y se caracteriza la colección obtenida en la sección 3. A continuación, se presentan las características de los sitios y de las páginas y en la sección 5, la topología del grafo web. Finalmente, se concluye y se exponen recomendaciones acerca de las debilidades halladas.

2 – Recolección de Datos

Para este trabajo se realizó una tarea de recolección de páginas web sobre el dominio estudiado utilizando el software *crawler* WIRE [Castillo, 2005]. La misma se efectuó en el mes de julio de 2006 utilizando un equipo servidor con un procesador de 2.4 GHz y 1 GB de memoria RAM corriendo el sistema operativo Linux Debian Sarge y un enlace a Internet de 2 Mbps.

Se recolectaron únicamente las páginas bajo el dominio ".py". El *crawler* fue inicialmente alimentado con aproximadamente 4.500 direcciones de dominios de segundo nivel provistas por el NIC Paraguay[3] a través de su sistema de consulta. El módulo de *crawling* se configuró con los siguientes parámetros: profundidad máxima en páginas dinámicas: 5 y profundidad máxima en páginas estáticas: 15. Por otro lado, se limitó la recolección sólo a páginas HTML, no incluyendo imágenes y otros objetos multimedia, y se descargó hasta un límite de 100 KB por página y hasta 40.000 páginas web por sitio.

3 – La colección WebPY

En total, se recolectaron 214.548 páginas desde 2.812 sitios que corresponden a 2.610 dominios de tercer nivel. Esta muestra se completó cuando el módulo recolector finalizó su tarea al no encontrar más enlaces por explorar. Si bien se considera este número como una cota inferior en el tamaño de la web de paraguay, pueden existir nuevos sitios o dominios que aún no se encuentran conectados con esta porción.

En cuanto a la cantidad de páginas, el 97,10% corresponden a únicas y el 2,90% se encuentran duplicadas. Del total, el 63,14% son páginas estáticas, mientras que las dinámicas suman el 36,86%. La tecnología más utilizada para la construcción de sitios dinámicos es PHP dado que el 84,35% de las dinámicas corresponden a este lenguaje. Según datos extraídos de NIC Paraguay la distribución de los dominios de segundo nivel es la que se presenta en la tabla 1:

	Registrados en NIC PY (julio 2006)		**Dominios activos**		**Proporción entre registrados y activos**
Dominio de 2do nivel	**Cantidad**	**%**	**Cantidad**	**%**	
com.py	4149	86,87	2151	82,41	0,52
org.py	374	7,83	260	9,96	0,70
net.py	39	0,82	22	0,84	0,56
gov.py	98	2,05	78	2,99	0,80
mil.py	1	0,02	1	0,04	1,00
edu.py	115	2,41	80	3,07	0,70
Otros	0	0	18	0,69	-
Total	4776		2610		

Tabla 1: Distribución de dominios de segundo nivel
Fuente: NIC Paraguay, julio 2006

En esta tabla se incluyen los dominios activos detectados luego de la etapa de recolección de datos. Aquí, consideramos como dominio activo a aquellos que respondieron exitosamente al menos una vez por recursos. Si observamos las proporciones entre registrados y activos encontramos que para el subdominio com.py hay una alta porción de dominios inactivos, especialmente si se tiene en cuenta que – en la actualidad – existe un costo asociado a la registración. Una explicación posible puede estar en que haya muchos dominios registrados antes de que exista el cargo de registración.

[3] http://www.nic.py

Inicialmente, realizamos una distribución de los principales códigos de error entregados por los servidores web al *crawler* al intentar descargar cada recurso. Dichos códigos fueron divididos en las siguientes categorías:

OK	Peticiones con estado de éxito
MOVED	Códigos que indican que el servidor redirige la petición a otra URL alternativa
SERVER ERROR	Peticiones que arrojan por resultado una falla producida del lado del servidor
FORBIDDEN	Peticiones que son denegadas por el servidor.

En la figura 1 se muestra la distribución de páginas de acuerdo a los grupos de códigos de estado propuestos. Aquí, observamos un 85% de transferencias exitosas lo cual demuestra que no hubo problemas mayores de conectividad, ni de disponibilidad de servicios. Por otro lado, los vínculos o enlaces rotos están en el 5,5% lo cual indica valores normales comparado con otros espacios web [Baeza-Yates, 2005_d]. Sin embargo, este valor indica algún problema de mantenimiento de los sitios, especialmente si se tiene en cuenta la existencia de herramientas y servicios que automatizan la tarea.

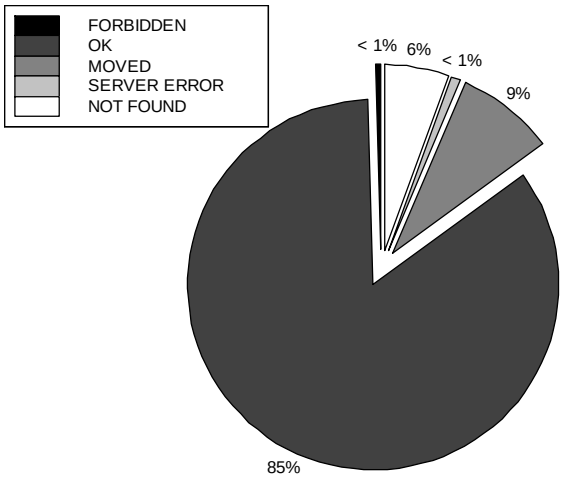

Figura 1 – Distribución de los códigos de estado

4 – Características de los Sitios y de las Páginas Web

Se estudiaron una serie de parámetros de los sitios y de las páginas web a los efectos de determinar características generales del espacio web de Paraguay. Como se mencionó, se recolectaron 214.548 páginas desde 2.812 sitios. Si comparamos estos valores con estudios sobre otras webs de la región, encontramos más de 3 millones de páginas en Chile [Baeza-Yates, 2005_a] y más de 7 millones Brasil [Modesto, 2005]. A nivel de sitios, se obtuvieron datos sobre cantidad promedio, edad y enlaces de sus páginas. Tales observaciones – entre otras – se presentan en la tabla 2.

Aquí también se puede apreciar una cantidad de sitios muy inferior a la observada en

otros países. Por ejemplo, en Chile se observaron más de 53.000 sitios y en Brasil más de 132.000. En cuanto al promedio de páginas por sitio encontramos 79 en Paraguay, valor similar a Brasil (85) y superior a Chile (57).

Sitios que aportaron recursos	2812
Sitios sin enlaces entrantes	1770
Sitios sin enlaces salientes	1995
Sitios con edad correcta en páginas	2249
Promedio de páginas por sitio	79,57
Promedio de páginas estáticas por sitio	48,42
Promedio de dinámicas por sitio	31,15
Promedio de edad de la página más antigua (meses)	17,46
Promedio de edad de la página más nueva (meses)	12,78
Promedio del tamaño de sitios (en MB)	1,49
Promedio de profundidad de sitios	2,78

Tabla 2 – Información resumen de sitios

4.1 – Tamaño de los sitios

El tamaño de los sitios se calculó en dos plano: cantidad de información que poseen (tamaño del sitio en bytes) y cantidad de documentos. En la tabla 3 se presentan los primeros veinte sitios de mayor volumen.

	Ranking por tamaño del sitio		**Ranking por cantidad de documentos**	
Orden	**Sitio**	**Tamaño (MB)**	**Sitio**	**Documentos**
1	Foros.hispavista.com.py	906	foros.hispavista.com.py	32.073
2	guia.hispavista.com.py	865	guia.hispavista.com.py	22.379
3	www.e2k.com.py	501	www.e2k.com.py	15.000
4	foro.juegosonline.com.py	354	foro.juegosonline.com.py	15.000
5	www.cdfenix.com.py	186	interno.roshka.com.py	6.717
6	www.leyes.com.py	108	www.leyes.com.py	4.808
7	www.laws.com.py	101	www.laws.com.py	4.483
8	www.embratel.com.py	84	www.cdfenix.com.py	3.793
9	www.guiacomercial.com.py	70	www.embratel.com.py	3.122
10	www.mercadeo.com.py	70	www.mercadeo.com.py	2.858
11	www.itacom.com.py	48	www.itacom.com.py	2.701
12	www.bloomberg.com.py	46	correo.aventuracard.com.py	2.607
13	www.holcim.com.py	39	www.guiacomercial.com.py	2.440
14	interno.roshka.com.py	39	www.bcp.gov.py	2.171
15	www.mug.org.py	33	www.lanacion.com.py	1.950
16	www.bcp.gov.py	32	www.jldeportivo.com.py	1.911
17	www.quanta.net.py	32	www.celular.com.py	1.881
18	www.celular.com.py	28	www.apg.org.py	1.725
19	www.arp.org.py	23	www.bloomberg.com.py	1.658
20	www.mobil.com.py	20	www.opusdei.org.py	1.642

Tabla 3 – Primeros veinte sitios con mayor cantidad de contenidos

En otros estudios sobre espacios web [Baeza-Yates, 2005_a] [Modesto, 2005] se

encontró que estas distribuciones siguen leyes de potencia [Adamic, 2002], es decir, existen pocos sitios que poseen altos valores (en este caso, bytes y documentos), mientras que hay muchos con valores bajos. En la web de Paraguay se verificó que las distribuciones de los tamaños de sitios – en bytes y documentos – siguen leyes de potencias.

Una cuestión a marcar es la ausencia de sitios del dominio educativo en los primero lugares de ambos rankings. Si bien la tabla 3 solo muestra las 20 primeras posiciones, como estas distribuciones siguen leyes de potencias, los sitios restantes son pequeños y muy alejados de los de los primeros puestos.

4.2 – Edad de las páginas

La edad de las páginas se determinó consultando la fecha de última modificación entregada por los servidores web en cada requerimiento. Este atributo es de suma importancia dado que habla acerca de la dinámica y mantenimiento del espacio. En algunos casos, la fecha devuelta fue errónea (servidores que no están con mantenimiento adecuado y su reloj interno se ha desfasado). La tabla 4 muestra la distribución de las edades de los documentos. Aquí, observamos que el 60% de las páginas se crearon o actualizaron en el último año.

Por otro lado, la figura 2 muestra el detalle de la fracción de páginas correspondiente al último año. Aquí hay que tener en cuenta que las indicadas con tiempo 0 incluyen las páginas dinámicas, las cuales se generan en el momento que se solicitan.

Edad en meses	Documentos	Fracción
Entre 0 y 12 meses	37.783	0,600
Entre 13 y 24 meses	8.885	0,141
Entre 25 y 36 meses	4.325	0,069
Entre 37 y 48 meses	6.416	0,102
Entre 49 y 60 meses	3.557	0,057
Entre 61 y 72 meses	836	0,013
Más de 72 meses	1.120	0,018

Tabla 4 – Distribución de las edades de los documentos

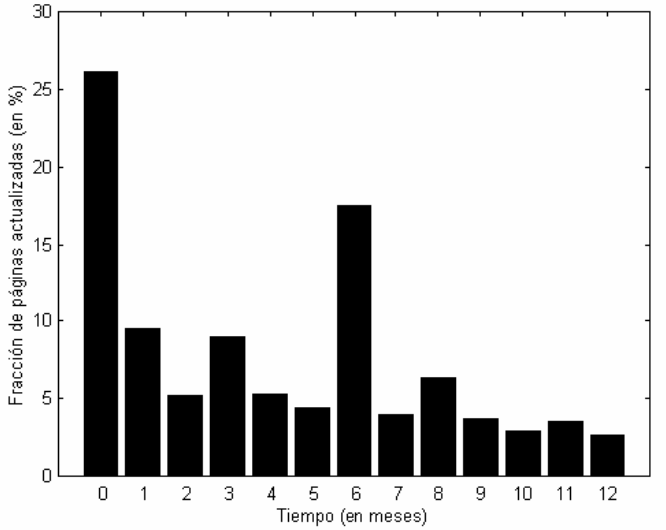

Figura 2 – Fracción de páginas actualizadas el último año

4.3 – Profundidad máxima de los sitios

La profundidad lógica de una página se define como el número de enlaces que es necesario seguir desde la página principal del sitio (profundidad 0) hasta ésta. La distribución de páginas por profundidad se muestra en la figura 3.

Como se mencionó anteriormente, el software se configuró para recolectar hasta la

profundidad 5 para documentos dinámicos y 15 para estáticos. Las distribuciones halladas son similares a [Baeza-Yates, 2005a] [Modesto, 2005]. Esta característica es importante a los efectos de ajustar eficientemente los parámetros de cualquier software de recolección.

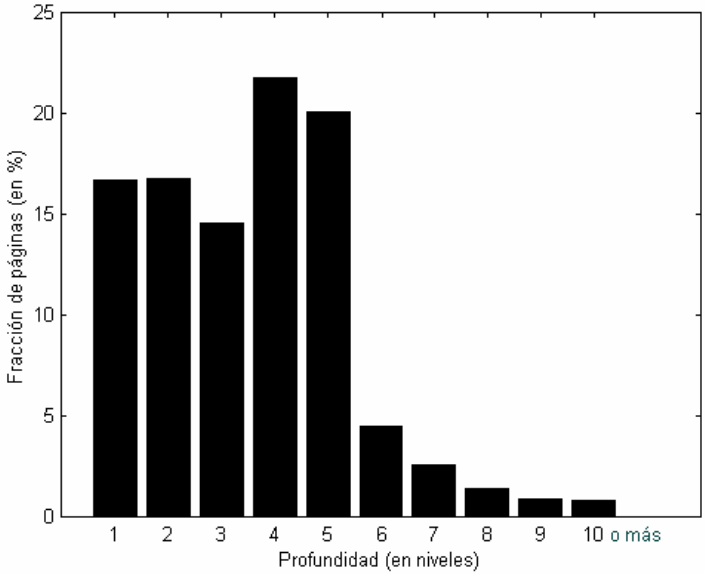

Figura 3 – Profundidad de los sitios

4.4 – Páginas por sitio

De los 2.812 sitios que al menos respondieron a los requerimientos del módulo de recolección, encontramos 1.112 sitios que poseen una sola página y pocos sitios grandes (Tabla 5). El valor más llamativo (y preocupante) es el primero, ya que resulta un indicio más de la necesidad de desarrollo de este espacio web.

Rango	Sitios
1 página	1.112
Entre 2 y 10 páginas	617
Entre 11 y 50 páginas	743
Entre 51 y 500 páginas	304
Entre 501 y 5000 páginas	32
Más de 5000 páginas	5

Tabla 5 – Distribución de la cantidad de páginas por sitio

5 – Topología

Se estudió este espacio web a partir de sus propiedades como grafo (*webgraph*) [Broder, 2000]. En éste, los nodos corresponden a páginas HTML y los enlaces entre éstas son las aristas. Este grafo forma una red libre de escala [Barabasi, 1999] en la cual porciones menores de ésta mantienen propiedades de la red completa (autosimilar) [Dill, 2002].

5.1 – Grado entrante

El grado entrante de una página corresponde al número de enlaces que apuntan hacia ésta

(en este caso desde el dominio estudiado). Es un indicador de reputación y es generalmente utilizado por los motores de búsqueda para el ranking de sus salidas [Page, 1998]. Las distribuciones son muy sesgadas y siguen leyes de potencias. La tabla 6 muestra los 20 primeros sitios de mayor grado entrante.

Orden	Sitios	Grado entrante (in degree)
1	www.abc.com.py	131
2	www.lanacion.com.py	68
3	www.senatur.gov.py	44
4	www.ultimahora.com.py	42
5	www.uninet.com.py	41
6	www.tigo.com.py	40
7	www.bcp.gov.py	40
8	www.presidencia.gov.py	39
9	www.telecel.com.py	36
10	www.rieder.net.py	34
11	www.webdesign.com.py	32
12	www.vox.com.py	32
13	www.mic.gov.py	31
14	www.hacienda.gov.py	30
15	www.dgeec.gov.py	29
16	www.pla.net.py	28
17	www.bestpisos.com.py	28
18	www.una.py	26
19	www.riederusados.com.py	23
20	www.lacueva.com.py	22

Tabla 6 – Distribución de sitios por grado entrante

Como dato relevante, encontramos el 41,6% de los sitios con grado entrante en cero. Teniendo en cuenta la importancia para el ranking, esto se puede ver como una debilidad y se requiere mejorar la conectividad, especialmente, entre sitios del mismo dominio.

5.2 – Dominios referenciados

En la tabla 7 se muestra la distribución de enlaces hallados en las páginas de la colección WebPY, agrupados por dominios de primer nivel. Se puede observar que el 93,8% de los enlaces corresponden a sitios externos a Paraguay.

Dominio de primer nivel	Número de enlaces externos	Fracción
com	4.813.861	0,701
py	1.624.549	0,237
es	206.481	0,03
net	54.145	0,008
org	34.374	0,005
us	31.116	0,005
ar	11.189	0,002

br	10.664	0,002
cl	9.748	0,001
cu	7.364	0,001
Otros	59.176	0,008

Tabla 7 – Primeros 10 dominios más referenciados

5.3 – Rankings de sitios

En este estudio se presenta un ranking denominado Siterank (tabla 8) basado en el algoritmo Pagerank [Page, 1998] donde se obtiene para cada sitio un puntaje, que indica la importancia del mismo basado en la cantidad de enlaces externos que recibe ponderados por la importancia de cada sitio emisor. Otro ranking utilizado es el basado en el algoritmo HITS [Kleinberg, 1999] donde se obtiene para cada sitio dos puntajes: Hub y Authority (tabla 8). El valor Hub indica qué tan bueno es el sitio como recurso de enlaces, en términos de qué tan importantes son los enlaces que tiene hacia otros sitios. En cambio, el valor Authority indica qué tan bueno es el sitio como recurso de información, en términos de qué tan importantes son los enlaces que recibe.

Orden	Ranking: Hub	Ranking: Authority	Ranking: Siterank
1	www.laws.com.py	www.alianza.com.py	guia.hispavista.com.py
2	www.leyes.com.py	www.localhost.com.py	foros.hispavista.com.py
3	www.cdfenix.com.py	www.nosotros.com.py	www.e2k.com.py
4	www.subastas.com.py	www.abc.com.py	foro.juegosonline.com.py
5	www.arp.org.py	www.leyes.com.py	interno.roshka.com.py
6	www.compraventa.com.py	www.conexion.com.py	www.leyes.com.py
7	www.lanacion.com.py	www.presidencia.gov.py	www.embratel.com.py
8	www.par.net.py	www.itacom.com.py	www.itacom.com.py
9	www.agronegocios.com.py	www.uninet.com.py	www.laws.com.py
10	www.elrodeo.com.py	www.adsl.net.py	www.guiacomercial.com.py
11	www.rafaelfilizzola.com.py	www.guayaki.com.py	www.opusdei.org.py
12	www.mre.gov.py	www.ultimahora.com.py	www.bloomberg.com.py
13	www.vue.org.py	www.lanacion.com.py	www.bcp.gov.py
14	www.paraguay.gov.py	www.webdesign.com.py	www.quanta.net.py
15	www.capadi.org.py	www.senatur.gov.py	www.mercadeo.com.py

Tabla 8 – Rankings de autoridades, hubs y siterank

5.4 – Macroestructura del espacio web

Analizado el espacio web como un grafo dirigido es posible caracterizar su dinámica. En él existe una componente fuertemente conexa, la cual es un subconjunto de los nodos del grafo donde existe un camino entre cualquier par de ellos. Las componentes fuertemente conexas que poseen más de un sitio no son demasiadas y la de mayor tamaño se llama “principal” o *core*. Según Broder [Broder, 2000] y Baeza-Yates [Baeza-Yates, 2005_a] se puede realizar una clasificación de los sitios de un espacio web de acuerdo a su relación con la componente fuertemente conexa principal:

MAIN:	Componente fuertemente conexa principal
MAIN-MAIN:	Sitios relacionados directamente con IN y con OUT.

MAIN-IN:	Sitios relacionados directamente con IN, pero no con OUT.
MAIN-OUT:	Sitios relacionados directamente con OUT, pero no con IN.
MAIN-NORM:	Sitios en MAIN que no corresponden a ninguna de las categorías vistas.
IN:	Sitios que llegan a MAIN, pero de MAIN no se puede llegar a ellos
OUT:	Sitios a los que se llega desde MAIN, pero no se puede retornar
TUNNEL:	Sitios en caminos de IN a OUT sin atravesar MAIN.
TENTACLE:	Sitios a los que se llega de IN o van a OUT, y no están en MAIN ni en TUNNEL.
ISLANDS:	Sitios no conectados a nada de lo anterior

Componente	Número de sitios	Fracción
MAIN	312	0,11
MAIN NORM	93	
MAIN MAIN	94	
MAIN IN	56	
MAIN OUT	69	
IN	278	0,10
OUT	479	0,17
Tentacle IN	99	0,04
Tentacle OUT	79	0,03
Tunnel	8	0,00
Islands	1.557	0,55

Tabla 9 – Macroestructura del espacio web de Paraguay

Como se puede apreciar en la tabla 9, la web estudiada se encuentra débilmente interconectada. Esto se debe a que el 55% de los sitios se encuentran en islas y – además – las regiones MAIN e IN son pequeñas (aproximadamente, 10%). La siguiente componente en tamaño es OUT (17%), pero – generalmente – corresponde a las páginas más antiguas, las cuales ya no son mantenidas (en cuanto a enlaces).

6 – Conclusiones

El espacio web de la República del Paraguay aún es un espacio reducido y requiere de mayor desarrollo para satisfacer – en forma mínima – las necesidades de información de sus usuarios. Posee más de 2.800 sitios que corresponden a más de 2.600 dominios activos. En tales sitios se recolectaron más de 210.000 páginas. Existe cerca de un 5,5% de enlaces rotos. El promedio de páginas por sitio es de alrededor de 79 documentos y el tamaño promedio del sitio es de 1,49 MB. El promedio de profundidad de los sitios es de 2,78 niveles. El 60% de las páginas de la web Paraguaya fue creada o actualizada en el último año, lo cual indica que se encuentra en crecimiento a una tasa alta respecto de su realidad.

Se hallaron 1.770 sitios sin enlaces entrantes, lo cual habla acerca de una falta de interconectividad a nivel nacional. Si se tiene en cuenta la importancia de los enlaces entrantes para el rankeo esta situación se puede entender como un problema importante. Según el análisis de macroestructura de Broder se observa que el 55% de los sitios son islas. El 93,8% de los enlaces externos son a dominios .com (70,01%) y .py (23,7%).

En cuanto a la herramienta de visualización utilizada para este trabajo, el concepto de *siteclouds* aportado permite hallar rápidamente los sitios más importantes de la web de Paraguay, de acuerdo a varios criterios. De esta manera, se puede realizar – también a a nivel de sitios – una exploración alternativa en búsqueda de los elementos más distinguibles.

Dadas algunas de las características encontradas sugerimos que se deberían aumentar los esfuerzos pare el desarrollo del espacio web de Paraguay. En especial, se requieren políticas que ayuden a incrementar la existencia, tamaño y posicionamiento de sitios relacionados con la educación. Complementariamente, se deben mejorar las estructuras de enlaces entre los sitios a los efectos de lograr mayor visibilidad. Todos estos puntos podrían ser parte de un proyecto de investigación e incorporación de tecnologías web a nivel nacional llevado adelante por sus Universidades.

7 – Referencias

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

[Baeza-Yates, 2003] R. Baeza Yates, B. Poblete y F. Saint-Jean, Evolución de la Web Chilena 2001-2002, Estudio Técnico, Centro de Investigación de la Web (CIW), 2003.

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

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

[Baeza-Yates, 2005_b] R. Baeza-Yates, C. Castillo and V. Lopez. Characteristics of the Web of Spain. Cybermetrics, Vol. 9, No. 1, 2005.

[Baeza-Yates, 2005_c] 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.

[Baeza-Yates, 2005_d] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of National Web Domains. Technical Report, Universitat Pompeu Fabra, July 2005.

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

[Broder, 2000] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, A. y J. Wiener. Graph Structure in the Web. Proc. 9th International World Wide Web Conference (WWW9)/Computer Networks, 33(1-6),2000, pp. 309-320.

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

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

[Efthimiadis, 2004] 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.

[EL PAIS, 2006] Diario El País. "Entrevista a David Sifry, fundador de Technorati" Diario El País, 1 de junio 2006, disponible en: http://www.elpais.es/articulo/portada/David/Sifry/fundador/Technorati/soy/editor/siglo/XXI/elpcibpor/20060601elpcibpor_1/Tes/

[Godwin-Jones, 2006] R. Godwin-Jones. EMERGING TECHNOLOGIES Tag Clouds in the Blogosphere: Electronic Literacy and social Networking. Language Learning & Technolog, Vol. 10, No. 2, pp. 8-15, 2006.

[Kleinberg, 1999] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632, 1999.

[Menezes, 2001] C. Menezes. “Desarrollo de la Sociedad de la Información en América Latina y el Caribe” División de la Sociedad de la Información, UNESCO-Montevideo, 2001.

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

[Page, 1998] L. Page, S. Brin, R. Motwani y T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[Redondo, 2004] M. Redondo Taemboury y A. Ferrari. La Sociedad de la Información en la Argentina, Presente y Perspectivas 2004-2006. Telefónica S.A., 2004.

[Unesco, 2005] Informe Mundial de la UNESCO “Hacia las Sociedades del Conocimiento”, 2005.