Artículo producto de la investigación

Caracterización de la web de Colombia

Arturo Eraso Torres[1]
Sixto Enrique Campaña Bastidas[2]

Resumen

Caracterizar la Web de Colombia es descubrir los elementos que hacen parte de la maraña mundial de Internet limitada por el dominio «*.co*». Es también identificar las partes que la constituyen: las páginas Web, los sitios y los dominios dentro del contexto nacional del ciberespacio Colombiano, manejado desde sus inicios por la Universidad de los Andes. Esto, con el propósito de conocer lo que es Colombia frente a Internet desde su composición Web. El estudio se enfocó desde un punto de vista tecnológico, descriptivo y analítico, el cual fue apoyado con herramientas informáticas (denominadas crawlers o robots) que permitieron adquirir información desde diferentes puntos de Internet. Se logró de esa manera que los datos recolectados permitieran caracterizar los elementos estudiados.

El proceso investigativo tomó como elemento importante para alcanzar la caracterización de la Web de Colombia una muestra de direcciones Web, a la cual se denominó «semilla». Esta muestra, con ayuda del *crawler*, permitió ampliar el rango de sitios y dominios en estudio, logrando tener una base estadística importante a la hora de establecer conclusiones y características significativas de la Web en estudio. Como producto adicional –complementario al *crawler* utilizado– se desarrolló un graficador de dominios con el cual fue posible tener una visión gráfica del objeto estudiado. Se espera que este documento –como complemento a la investigación desarrollada– sea un punto de partida para ahondar mucho más sobre la estructura de la Web en el país.

Palabras clave: Crawler, Web, Nutch, Dominios, Segmentos, Mallas.

1 Magister en Software Libre. Docente TC I.U. CESMAG. Pasto – Colombia. aeraso3@gmail.com

2 Magister en Software Libre. Docente Auxiliar UNAD – ECBTI. Pasto –Colombia. sixto.campana@unad.edu.co

Web Colombia Characterization

Abstract

Colombia characterize the Web is to discover the elements that are part of the global Internet tangle limited by the ". CO" is to identify the parts which are: web pages, sites or domains within the national context of cyberspace Colombian managed since its inception by the University of the Andes, all with the purpose of learning about what Colombia is facing Web Internet since its composition, the study focused from a technological point of view, descriptive and analytical, which was supported with tools that allow to acquire information from different points of the Internet, making the collected data allowing to characterize the elements studied, these tools are called crawlers or robots. The research process took as an important element to achieve the characterization of a sample Web Colombia web addresses, to which was called "seed", which helps the crawler broadened the range of sites and domains in the studio, managing to have a important statistical basis which permitted conclusions and significant features of the Web study was developed as an additional product plotter domains, which is complementary to the crawler used and allowed to have a graphical view of the object being studied. This document is expected to complement the research undertaken as a starting point to delve much more about the Web structure of the country.

Key words: Crawler, Web, Nutch, Domains, Segments, grids.

Recibido: 10 de febrero 2012
Aceptado: 23 de marzo 2012

Introducción

El auge de Internet y todos los servicios que esta red contiene, han llevado al uso cotidiano y cada día más necesario, de términos como Web, Mail, entre otros. . Pero este crecimiento ha sido posible gracias a la organización mundial de la red, una organización que para muchos es invisible, pero que existe. Es el caso de la Web de cada país, la cual se encuentra caracterizada y abarca una gran variedad de aspectos como las políticas de gobierno, las capacidades tecnológicas, los conocimientos específicos, entre otros.; Contextos importantes

y relevantes de estudio, de los cuales se tiene referentes son los de Brasil (Modesto et al., 2005), España (Baeza-Yates *et al.*, 2006), Argentina (Tolosa *et al*., 2007) y Chile (Baeza-Yates y Castillo, 2005; Castillo et al., 2006). En Colombia se tiene antecedentes del investigador Guillermo Correa Uribe (1999), quien enfocó su análisis desde dos puntos de vista: uno técnico y otro visual o de observación personalizada. La mayor parte del proceso fue manual y con base en instrumentos de recolección de información validados mediante cuestionarios que se aplicaron a 2.976 sitios Web de los 11.864 registrados en 1998.

Teniendo en cuenta lo anterior, el presente estudio se convierte en la primera investigación de la Web de Colombia. Con la ayuda de herramientas de navegación automáticas denominadas *crawlers*, se han podido establecer unas características más específicas del ciberespacio Colombiano. Estas características van desde el tipo de páginas, su tamaño, lenguajes y documentos soportados hasta aspectos generales del colectivo (sitios) y terminando con algunas características específicas de los comúnmente denominados dominios de Internet.

La semilla: el punto de partida

El proceso de obtención de la información de la Web de Colombia se realizó mediante un *crawler*, el cual es un software que permite descargar el contenido de una lista de sitios predeterminada. A esta lista de sitios se le denominó semilla. Así pues, el *crawler* fue alimentado con aproximadamente 5200 direcciones Web (URLS) obtenidas a través de diversas fuentes (principalmente directorios en Internet) y tomadas como el punto de partida.

Acorde con lo anterior, se estableció la siguiente distribución de la semilla para la descarga de información de la Web de Colombia.

Tabla 1. Distribución de la semilla y porcentaje respecto a la Web Colombia

Nombre	Número	Porcentaje (%)
Com.co	2.124	11,06%
Gov.co	1.208	32,49%
Edu.co	1.207	52,9%
Org.co	524	42,36%
Net.co	99	35,7%
mil.co	37	29,3%
int.co	1	1,9%
arts.co	0	0,000%
info.co	0	0,000%

Como se puede observar en la Tabla 1, el dominio con mayor cantidad de sitios en la semilla es el comercial (*.com.co*) con 2.124 direcciones, que representan el 11,06% en relación con la Web de Colombia. En segundo lugar está el dominio gubernamental (*.gov.co*) con 1.208 direcciones, que forman el 32,49% de los sitios .gov.co en la Web de Colombia. En tercer lugar está el dominio de instituciones educativas (*.edu.co*) con 1.207 direcciones, que representan el 52,9% del total de sitios con esta denominación en la Web de Colombia.

En el cuarto lugar se tiene a las organizaciones no gubernamentales (*.org.co*) con 524 direcciones, que constituyen el 42,36% de sitios .org.co en la Web de Colombia. En el quinto lugar se tiene a las .net.co con 99 direcciones y una representación del 35,7% en relación con las direcciones del mismo tipo en la Web de Colombia. En el sexto lugar se tiene a las instituciones militares con 37 direcciones, que constituyen el 29,3% de las direcciones militares en la Web de Colombia. Finalmente el dominio perteneciente a internacional (*.int.co*) se encuentra presente con una dirección que constituye el 1,9% de las direcciones con esta denominación en la Web de Colombia. El total de direcciones de la semilla representa el 19,32% de sitios que constituyen la Web de Colombia y que a la postre se convirtió en la base para el estudio.

Descarga de la información

Un segundo elemento importante en la consecución de la información para el análisis de la Web de Colombia fue la herramienta destinada para descargar la información: el *crawler* Este elemento se analizó desde dos puntos de vista: características de la máquina desde la cual se debería ejecutar y configuración del mismo.

Características de la máquina

Los *crawler* son programas que se pueden considerar como no complejos en su instalación y ejecución pero que, dependiendo de la cantidad de información a descargar y de los parámetros que se configuren en los procesos a realizar, se pueden convertir en aplicativos bastante exigentes para la máquina que los soportará y que permitirá el cumplimiento del objetivo trazado. En este orden de ideas el programa que se instaló fue *Nutch* y el objetivo fue descargar la mayor cantidad posible de sitios de la Web de Colombia (a partir de 5.200 direcciones aproximadamente). Estas fueron las características de la máquina que permitió realizar dicho proceso: procesador Intel Pentium IV de 3 Ghz, memoria RAM de 2 Gb, 50 Gb de espacio libre en disco duro y sistema operativo Linux.

Configuración del crawler

El *crawler Nutch*, se configuró con el fin de capturar la mayor cantidad de información de los sitios de la Web de Colombia a partir de la semilla dada. Dicho procedimiento requirió la configuración de los siguientes elementos:

Profundidad: Se refiere a los niveles que debe avanzar el *crawler* con respecto a una dirección dada, es decir, hasta que punto se quiere que la dirección sea explorada. Cada sitio tiene diferentes páginas anidadas, así que al definir la profundidad se llegará a un mayor número de páginas del sitio propuesto. Para el estudio realizado, dicha configuración fue de 8 niveles como profundidad máxima.

Páginas de descarga por sitio: Este ítem se refiere al límite que se quiere tomar como parámetro al momento de explorar un sitio. Los dominios son conjuntos de múltiples páginas, algunas dinámicas y otras estáticas. Un sitio puede estar conformado por miles de páginas, por unas cuantas y, en algunos casos, por una sola. Tomando como punto de referencia estudios similares relacionados en el ítem anterior, se configuró un parámetro de 40.000–un valor manejable y que al mismo tiempo permite descargar datos necesarios para poder caracterizar el sitio en exploración. También este parámetro es el que permite que las direcciones aumenten, dado que si el *crawler* encuentra un nuevo vínculo, lo explorará y buscará más páginas del mismo.

Hay muchas más características que se podrían explicar y enumerar en la configuración del crawler, pero las más significativas son las relacionadas en los ítems anteriores.[3]

Características de las páginas web

Luego de haber realizado los procedimientos anteriores relacionados con la definición de la semilla, la elección del *crawler*, la selección de los mejores parámetros de configuración del mismo y el uso de una buena máquina para el desarrollo del proceso, se inició la descarga y con ello la obtención de los resultados. A continuación y durante toda esta sección se hablará de algunas de las características que se pudieron encontrar en las páginas que hacen parte de la Web de Colombia. Los ítems tratar son tamaño y tipos de páginas.

Tamaño de las páginas

En lo relacionado con el tamaño de las páginas se encontró que en la Web de Colombia el valor promedio es de 15,42 Kb. Este valor es menor que las caracterizaciones de la Web de Chile (21 kb) (Castillo *et al*., 2006), Brasil (24 kb) (Modesto *et al*., 2005) y mayor de la Web Argentina (10 kb) (Tolosa *et al*., 2007).

Tipos de páginas

Para el estudio de la Web de Colombia se dividió el conjunto de datos obtenido en dos grupos. Uno de ellos fue denominado «páginas dinámicas», y se refiere

3 Para conocer detalles sobre la configuración de Nutch el lector podrá remitirse a la dirección http://lucene.apache.org/nutch

a aquellas páginas que son generadas bajo petición, son capaces de responder de manera inteligente a las demandas de un cliente y permiten automatizar gran cantidad de tareas.

El otro grupo es el de las «páginas estáticas», que son las más comunes y simples: aquellas en las cuales su contenido poco o nada cambia en el tiempo.

Tabla 2. Distribución de los documentos estáticos y dinámicos

	Documentos	**Porcentaje**
Total	566.961	100
Estáticas	187.267	33,03
Dinámicas	379.694	66,97

En el análisis de datos expuesto en la Tabla 2 se puede observar que la mayoría de páginas son de tipo dinámico (66,97%), mientras el 33,03% corresponde a páginas estáticas. La Web Argentina reportó el 52% (Tolosa *et al*., 2007)de páginas dinámicas, en contraste España el 22% (Baeza-Yates *et al*., 2006) y Chile el 42,5% (Castillo *et al*., 2006). El porcentaje de la Web de Colombia de páginas dinámicas es bastante elevado. Esto denota que existe una importante infraestructura de desarrollo Web que soporta gran parte de la lógica de negocios de las organizaciones. Es una muestra, igualmente, de las políticas que ha implementado el Estado Colombiano en el sector público principalmente.

Características de los sitios web

Algunos de los elementos que se analizan en esta sección son tamaño en MB de un sitio y cantidad de enlaces.

Tamaño en Mb de un sitio

El tamaño se calculó con base al número de bytes por sitio. En la Tabla 3 se presentan los 10 sitios con mayor tamaño en la Web de Colombia. Se puede observar que el mayor porcentaje pertenece a instituciones educativas, seguidas de entidades estatales y organizaciones no gubernamentales.

Tabla 3. Primeros diez sitios con mayor tamaño

Orden	**Sitio**	**Tamaño**
1	industrial.edu.co	711,37
2	redacademica.edu.co	689,73
3	scc.org.co	273,81

Orden	Sitio	Tamaño
4	unisabna.edu.co	243,02
5	uis.edu.co	169,34
6	unal.edu.co	162,83
7	gi.edu.co	161,80
8	une.net.co	140,83
9	cnrr.org.co	130,67
10	dane.gov.co	126,85

Cantidad de enlaces

Este apartado hace referencia al estudio de las relaciones establecidas a nivel de enlaces. El espacio web es modelado como un grafo dirigido sobre el cual se analizan diferentes características.

Grado entrante

El número de enlaces que recibe una página Web se denomina grado entrante (*in-degree*). La información recolectada con el *crawler* mostró que la Web de Colombia a partir de la semilla aplicada tiene 60.091 enlaces entrantes. La Tabla 4 muestra los 10 primeros sitios de mayor grado entrante.

Tabla 4. Primeros diez sitios con mayor grado de enlaces entrantes

Orden	Sitio	Grado Entrante
1	unal.edu.co	2.416
2	quebarato.com.co	1.882
3	dnp.gov.co	1.213
4	unalmed.edu.co	1.160
5	unisabana.edu.co	1.043
6	competitividad.gov.co	760
7	cordoba.gov.co	676
8	camaramedellin.gov.co	664
9	interactic.com.co	655
10	snc.gov.co	594

Se encontró que 3,26% de los enlaces entrantes tienen un valor igual a cero. El dominio que más enlaces entrantes registró fue *.gov* con 22.288 y el que menos tuvo fue el *.mil* con 512.

Grado saliente

El grado saliente se refiere a la cantidad de enlaces que posee una página con respecto a una externa (*out-degree*). Para el caso de la Web de Colombia a partir de la semilla aplicada y utilizada por el *crawler*, se encontraron 74.707 enlaces salientes. La mayor participación nuevamente la tuvo *.gov* con 28.482 enlaces y la menor *.mil* con 367. En la Tabla 5 se puede ver los 10 sitios con grado saliente más relevante:

Tabla 5. Primeros diez sitios con mayor grado de enlaces salientes

Orden	Sitio	Grado entrante
1	Gobiernoenlinea.gov.co	3.203
2	cancilleria.gov.co	2.592
3	presidencia.gov.co	1.907
4	nuevoestadio.com.co	1.600
5	buscape.com.co	1.514
6	unal.edu.co	1176
7	santillana.com.co	1.144
8	contratos.gov.co	1.122
9	universia.net.co	1.113
10	servientrega.com.co	929

Características de los dominios

Por último, la descripción de las características encontradas en la Web de Colombia corresponden a los dominios, de los cuales se tuvo en cuenta los siguientes ítems: número de sitios por dominio y la estructura macroscópica.

Número de sitios por dominio

En total existen 5.971 sitios que corresponden a 3.240 dominios. La mayoría de los dominios de la Web de Colombia contiene un solo sitio, lo que quiere decir que la administración no proporciona los espacios de organización necesaria para tener una correspondencia de datos o que la cantidad de información no amerita una jerarquía organizativa. Esto se puede observar en la Tabla 6.

Tabla 6. Distribución de sitios por dominio

	Cantidad	Porcentaje (%)
Total de sitios en la muestra	5.971	
Total de dominios	3.240	

	Cantidad	Porcentaje (%)
Dominios con mas de un sitio	492	15,19
Dominios con un solo sitio	2.748	84,81

El número de sitios por dominio se observa en la Tabla 7. Se trató de guardar una referencia entre la cantidad de dominios suministrados por la NIC Colombia y la semilla, obteniendo una proporción de crecimiento en la mayoría de dominios. Por tal motivo se obtuvo una mayor cantidad de datos en el dominio com.co (36,36%).

Tabla 7. Número de sitios por dominio

Dominios	No de sitios por dominio
Com.co	2.171
Edu.co	1.902
Gov.co	1.180
Org.co	576
Net.co	100
mil.co	40
int.co	1

Dominios genéricos

En este apartado se ha realizado una descripción a nivel de dominios genéricos de la Web de Colombia, basados específicamente en los *com.co, edu.co, gov.co, mil.co, net.co y org.co.* Se usaron datos como la cantidad de dominios, sitios, links internos, enlaces entrantes, enlaces salientes, total páginas descargadas y tamaño.

Dominio com.co

Es el dominio con mayor tamaño de la Web de Colombia (45,69%) y es a la vez el que más sitios posee (36,36%). Cuenta con el mayor número de enlaces internos (40,99%), sin embargo cuenta con el tercer porcentaje de enlaces entrantes (24,98%) y el segundo de porcentajes salientes (33,13%),lo que supone falta de conectividad entre las páginas para poder llegar más efectivamente de un punto a otro. Presenta el mayor número de páginas (45,27%) pero no de mayor tamaño de información (14,99%).

Dominio edu.co

El dominio *edu.co* representa el 13,92% de la Web de Colombia, contiene 1.902 sitios (31,86%), cuenta con el 646.811 enlaces internos (27,87%) y tiene el segundo porcentaje de enlaces entrantes (26,22%) y el tercero de

porcentajes salientes (15,74%). Esto supone una falta de interconexión entre las instituciones educativas. De igual forma, este dominio presenta el segundo porcentaje en número de páginas con el 27,19% y el de mayor tamaño de información (43,05%).

Dominio gov.co

El dominio *gov.co* es el que representa mayor interacción entre sitios entrantes y salientes: 37,09% y 38,12% respectivamente. Posee el tercer número de sitios recolectados (19,77%) y representa el 25,22% de la Web de Colombia con 817 dominios. Posee el segundo mayor tamaño con el 25,06% y el tercer porcentaje de páginas con el 16,47%.

Se puede concluir entonces, que los dominios colombianos que presentan un mejor desarrollo actualmente son *edu.co y gov.co*, lo que da a conocer el soporte de las políticas del Estado por parte de sus diferentes instituciones y la implementación de sus procesos en la Web. Las instituciones educativas por su parte se deben ver beneficiadas por este tipo de tecnologías al maximizar sus recursos en el manejo de las TIC. El dominio *com.co* no cuenta con un proceso de desarrollo acorde que facilite su integración ya que, como fue posible evidenciar en el estudio, existen páginas de dominios cerradas y sin posibilidad de interconexión hacia otros lugares.

Estructura macroscópica

La Web de Colombia se encuentra débilmente interconectada. Esto se debe a que el MAIN solo alcanza el 25,94%, en comparación con la Web Argentina que posee el 54,23% (Tolosa *et al*., 2007). Quizá se puede ver valores más cercanos al caso colombiano en la Web chilena con un 21,76% (Castillo *et al*., 2006)y en Brasil que alcanza el 25,27% (Modesto et al., 2005) . Los sitios que componen el OUT representan el 27,59%, que es un valor similar al de Chile donde representan un 26,12% (Castillo *et al*., 2006) pero inferior a Brasil con el 45,33% (Modesto *et al*., 2005).

Los sitios correspondientes a IN e ISLANDS se los accede a partir de sus páginas iniciales debido a que pueden ser páginas nuevas o no estar bien conectadas. En la Web de Colombia estos valores son 11,42% para IN y 29,39% para ISLANDS, en Chile 6,65% para el componente IN y 46,16% en ISLANDS (Castillo *et al*., 2006) y en Brasil 12,95% y 12,35% respectivamente (Modesto *et al*., 2005).

Grafos de la web Colombia

Como un producto adicional se desarrolló un script que permitió obtener de manera gráfica la estructura de los diferentes dominios que hacen parte de

la Web de Colombia. En este apartado se presentan de manera gráfica los resultados obtenidos a partir de los dominios genéricos de primer nivel *com.co, gov.co y edu.co*. Se podrá observar principalmente su distribución y los enlaces entrantes como salientes.

Grafo com.co

Como se puede observar en la figura 1 del gráfico del dominio com.co, cada punto rojo identifica un dominio. Se presenta aquí una gran cantidad de puntos rojos aislados y sin conexión, lo que significa que la maraña o red se corta demasiado., También es una explicación al porqué no crece la cantidad de sitios con la semilla propuesta: son conexiones independientes que no permiten ir a otros sitios dentro del mismo dominio. Lo anterior, indica, en pocas palabras, que el dominio .com.co no está fuertemente conectado.

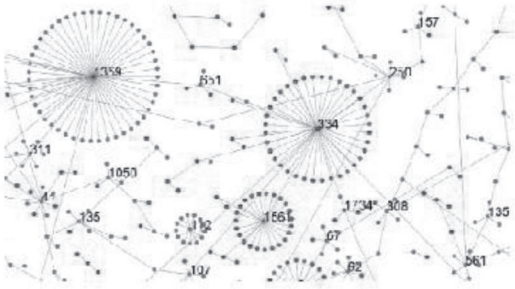

Figura 1. Sección gráfico .com.co

Grafo gov.co

De la misma manera como se hizo con el dominio *.com.co*, se ha obtenido el gráfico del dominio *.gov.co.* Este dominio generó un gráfico bastante complejo de entender, por ello se ha tomado y analizado una muestra(Figura 2).

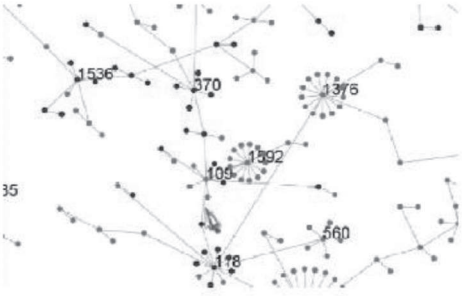

Figura 2. Sección gráfico .gov.co

El color azul es el que identifica los nodos *.gov*. En el grafo se puede observar que a diferencia del dominio *.com.co*, aquel se encuentra más interconectado entre sí y son pocos los nodos que se encuentran aislados. Esto significa que el dominio *.gov.co* esta fuertemente conectado y de un nodo se puede llegar a otros dentro del mismo dominio –aunque la tendencia sigue siendo que la mayoría son enlaces externos.

Conclusiones

La distribución del idioma está compuesta por un porcentaje del 96,78% en español seguido del inglés con el 3,22 %.

El 90% de las páginas han sido creadas o modificadas en el último año, lo que demuestra que el espacio Web colombiano está creciendo como ocurre en otros países.

En cuanto a los aspectos tecnológicos, el 33,03% del total de las páginas descargadas son de tipo estático y el 66,97% dinámico. Dichas páginas se encuentran construidas en gran parte por herramientas de tipo libre como PHP con el 48,33% y SHTML con 43,25%.

La Web en Colombia se encuentra débilmente interconectada. Un indicador es que el componente MAIN solo alcanza el 25,94% (debajo de los indicadores de otras Web de estudio). Sumado a esto, existe un alto porcentaje correspondiente a IN e ISLANDS:11,42% y 29,39% respectivamente.

Los dominios que presentan mayor desarrollo en la Web Colombia por contenido y conexión son el gov.co y el edu.co, lo cual exige una mayor atención a las estructuras de los enlaces entre los sitios.

La Web de Colombia ha presentado un crecimiento de más del 800% desde su último estudio que data de 1998. Esto demuestra que Colombia no ha sido la excepción en el desarrollo y aceptación de Internet en el mundo. De hecho, el país ha guiado su desarrollo tecnológico a través del uso de esta alternativa de comunicación e información.

Colombia y su Web crecerán mucho más. Las políticas gubernamentales han demostrado que es posible mantener una estructura sólida en comunicación tecnológica; o por lo menos los resultados obtenidos del estudio del dominio .gov.co así lo demuestran. En este mismo camino se encuentran las entidades educativas, pero el sector comercial tendrá que trabajar mucho más para lograr una fuerte intercomunicación de sus dominios y así mostrar que la Web de Colombia está fuertemente conectada y, que se puede ampliar su cobertura y alcance.

Referencias Bibliográficas

Baeza-Yates, R. y C. Castillo. 2005. *Características de la web chilena* 2004. Santiago de Chile: Centro de Investigaciones de la Web.

Baeza-Yates, R. *et al*. 2006. «Características de la web de España». *El Profesional de la Información* 15(1): 6-17.

Castillo, C. *et al*. 2006. *Características de la Web chilena.* Santiago de Chile: Centro de Investigaciones de la Web.

Correa Uribe, G. 1999. *Colombia conectada al mundo: sitios web colombianos.* Medellín: Escuela Interamericana de Bibliotecología.

Modesto, M. *et al*. 2005. «Um novo retrato da web brasileira». *Proceedings of XXXII SEMISH: 2005–2017.*
Tolosa, G. et al. 2007. *Caracterización del espacio web de Argentina.* Buenos Aires: Universidad Nacional de Luján.