Análisis de la Web Oculta en España

Manuel Álvarez, Fidel Cacheda, Rafael López-García, Víctor M. Prieto.
Departamento de Tecnologías de la Información y las Comunicaciones,
Universidade da Coruña
Facultade de Informática, Campus de Elviña, S/N, 15071, A Coruña (Spain).
mad@udc.es, fidel@udc.es, rafael.lopez@udc.es, victor.prieto@udc.es.

***Resumen*- Este artículo presenta un estudio sobre los sitios web de los dominios ".es" orientado a determinar el nivel de utilización de determinadas tecnologías que dificultan el recorrido de la Web a los sistemas de *crawling*. En particular, el estudio se centra en dos aspectos relacionados con la "Web Oculta": los *scripts* y los formularios. En base a los resultados obtenidos, se concluye que un *crawler* que pretenda obtener la mayor parte de documentos de la Web debe de tratar tecnologías tales como *scripts* o formularios para conseguirlo.**

***Palabras Clave*-Recuperación de Información, Web Oculta, Web española, formulario, *script*.**

I. INTRODUCCIÓN

Se conoce como "Web Oculta" o "Web Profunda" [1] a la parte de la Web que no está directamente enlazada. Existe un conjunto de tecnologías que los sistemas de *crawling* convencionales no son capaces de tratar y que constituyen los puntos de acceso a esos documentos denominados "ocultos". Por una parte se pueden considerar los formularios web como puntos de entrada a la Web Oculta del lado servidor. Por otra parte, para acceder a la Web Oculta del lado cliente es necesario tratar con tecnologías como lenguajes de *scripting* o Flash.

Para determinar el nivel de utilización de estas tecnologías que dificultan el acceso a los documentos por los sistemas de *crawling*, en 2009 Álvarez et al. [2] comenzaron un estudio sobre los dominios ".es". El estudio se dividió en dos fases: (1) diseñar, implementar y ejecutar un "*crawler*" que descargase la primera página de dichos dominios a partir de una lista actualizada de los mismos y generase estadísticas cuantitativas y (2) analizar el contenido de las páginas para determinar las tecnologías que utilizan.

Este artículo aborda la segunda fase del estudio. Para ello, parte de la arquitectura definida en [2] y la extiende dotando al sistema de un analizador del contenido de las páginas.

La estructura de este artículo es la siguiente: en la sección II se repasan los trabajos relacionados con la materia. La sección III explica la arquitectura del *crawler* usado en el experimento. La sección IV analiza los resultados del experimento y en la sección V se explican las conclusiones y trabajos futuros.

II. TRABAJOS RELACIONADOS

La mayor parte de los estudios sobre la Web se ocupan de la Web de Superficie, aunque existen algunos que se ocupan de la Web Oculta [3]. Existen sitios web que ofrecen estadísticas sobre el contenido de la Web indexado [4], sobre el número de servidores web [5] o sobre el contenido de las páginas [6][7]. Por otra parte, existen organizaciones encargadas de mantener los nombres de dominios y de realizar el recuento de las máquinas dadas de alta en ellos [8]. Algunas, como Red.es [9], que es la que mantiene los dominios españoles, también publican datos sobre la evolución del número de dominios en el tiempo. Otras, como Verisign [10], que administra los dominios ".com" y ".net" hacen informes más completos. Sin embargo, no existen informes públicos que analicen las páginas de los sitios Web españoles para determinar las tecnologías que utilizan.

También son varios los trabajos que ponen de manifiesto las diversas dificultades con las que los *crawlers* tienen que lidiar para acceder a algunos documentos. Un ejemplo lo constituyen los lenguajes de *scripting* siguiendo el estándar ECMAScript [11]. Sin embargo, según reflejan Weideman y Schwenke en [12] y Wu y Davison en [13], aunque sean tecnologías ampliamente usadas, los *crawlers* no suelen evaluarlos en busca de URLs.

III. ARQUITECTURA

A diferencia de los *crawlers* convencionales, se ha implementado uno que no sigue los enlaces de las páginas web, sino que parte de una lista completa de dominios para obtener el estado de cada uno y el contenido de su página principal. Con estos datos, un módulo de análisis de *crawling* se encarga de generar las estadísticas de la fase 1. Para la fase 2 se ha añadido un módulo de análisis del contenido de los documentos, el cual almacena los datos de cada una de las páginas en una base de datos para facilitar la generación de estadísticas. La Fig. 1 muestra la arquitectura del sistema.

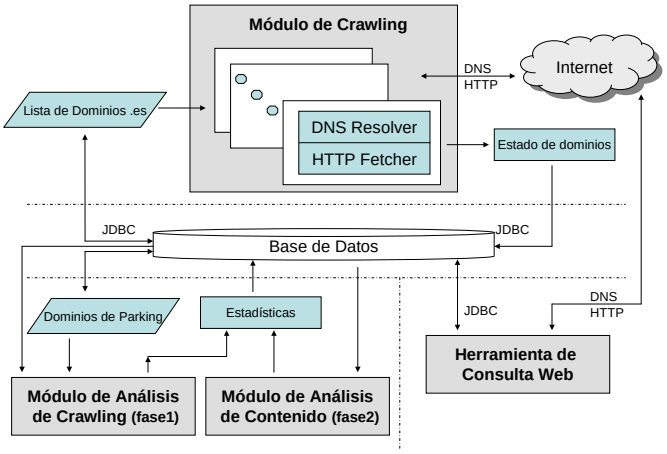

Fig. 1. Arquitectura del sistema de *crawling*.

El módulo de análisis de contenido emplea un analizador sintáctico XML que trata cada documento HTML como un recurso XHTML. Para ello usa el analizador CyberNeko

HTML [14]. El analizador XML obtiene la metainformación de la página e identifica diferentes elementos y tecnologías que pueden ser utilizados para navegar o generar contenido, como lenguajes de *scripting* o formularios.

IV. ANÁLISIS DE RESULTADOS

Según el estudio de Álvarez et al. [2], la Web española presentaba 1.093.193 dominios en mayo de 2009. De ellos, solo 577.442 (52,82%) tenían un servidor web. En los siguientes apartados se comentan los resultados obtenidos para el análisis del contenido de la página principal de los dominios ".es" con servidor web. Los resultados se muestran en base al nivel de utilización de lenguajes de *scripting* (IV.A), formularios (IV.B) y otras tecnologías (IV.C).

A. Scripts

Los *scripts* constituyen la principal barrera de acceso a la Web Oculta de lado cliente. En el caso de la Web española, se han encontrado *scripts* en 266.737 dominios, un 46,2% de los que tenían un servidor que no devolvía error. En cuanto al uso de ficheros de *script* externos, se han contabilizado 542.322 invocaciones en 179.576 dominios (31,1%), cifra inferior a los 744.111 *scripts* internos que se encontraron en sus respectivas etiquetas <script> en 231.059 dominios (40%). Ambas cifras son también inferiores a las del uso de *scripts* en atributos HTML, con 2.266.881 ocurrencias en 147.617 dominios (25,6%) .

Según el último RFC de "*Scripting Media Types*" [15] los *scripts* deberían indicar el lenguaje en el que están escritos. Sin embargo, la Tabla 1, basada en los 1.286.419 *scripts* que no se encontraban en atributos HTML, muestra que esto no siempre se cumple. Las invocaciones a scripts externos suelen indicarse de una forma un tanto más rigurosa, probablemente porque se añaden mediante programas de diseño web. Cuando estos mecanismos de identificación de lenguaje fallan, se puede determinar a través de la meta-información de la página, de un análisis del código o de las extensiones de los ficheros. Esta última no es formal, pero ofrece buenos resultados.

La gran mayoría de los *scripts* encontrados en la Web española siguen el estándar ECMAScript [11]. Es más, casi todos están escritos en lenguaje JavaScript.

Scripts	Internos	Externos	Total
Con "type"	75,30%	89,90%	87,70%
Con "language"	14,65%	6,95%	5,20%

Tabla 1. Identificación del lenguaje de los *scripts*.

Para el caso de *scripts* en atributos HTML, se ha comprobado que 137.802 dominios (un 23,8%) hacen uso de eventos "onXXX" sobre diferentes etiquetas. Sin embargo, solo 488.236 scripts en atributos HTML (un 21,5%) incluyeron la etiqueta "javascript:". Una cantidad residual (426) presentaba la etiqueta <script> en el atributo HTML.

También se ha realizado un estudio para determinar qué etiquetas de HTML suelen contener más código *script*. En general, un 37,5% de los dominios contienen bloques <script> en algún lugar dentro del <body>, mientras que solo un 25,7% de los dominios contienen dichos bloques fuera del mismo.

También se ha estudiado la localización de las etiquetas <script> según la etiqueta que las contiene, incluyendo tanto las llamadas a ficheros externos como los bloques de *script* embebidos. Para garantizar la visibilidad de los *scripts*, se recomienda que los bloques se incluyan en la etiqueta <head> o al principio del <body>. Sin embargo, muchos han aparecido en otras etiquetas. Un ejemplo es el de la etiqueta , que debería usarse para crear bloques de marcado. Los primeros resultados se pueden ver en la Tabla 2:

Etiqueta HTML	Número de scripts	Dominios
<head>	530.522	200.713
	260.275	75.619
<body>	228.887	99.361
<td>	116.191	43.992
<p>	31.488	13.941

Tabla 2. Localización de etiquetas <script> en HTML

La Tabla 3, por su parte, muestra las etiquetas que contienen más *scripts* en sus atributos:

Etiqueta HTML	Número de scripts	Dominios
<a>	1.305.831	95.402
<img>	196.419	18.880
<td>	184.657	6.495
	140.825	12.519
<input>	111.013	36.531

Tabla 3. Localización de scripts de eventos en HTML

Como se puede comprobar, la mayoría de los eventos están localizados en enlaces, en atributos "onXXX". En muchas ocasiones esto se hace para generar dinámicamente la URL a la que apuntan. Sin embargo, se han detectado *scripts* en el atributo "action" de los formularios, etc.

También se ha estudiado el evento "onLoad" de la etiqueta <body>, apareciendo en 47.064 dominios (un 8,2%).

Respecto al número de *scripts* empleados por página, tanto en general como separando las invocaciones a ficheros externos y los *scripts* embebidos (contando los que se alojan en atributos de la etiqueta HTML), se obtiene una distribución de ley de potencia con una cola larga. Esto quiere decir que la mayor parte de las páginas no invocan ningún *script*, un gran número invocan pocos *scripts* y solo unas pocas invocan un gran número de ellos. La Fig. 2 muestra las distribuciones del total de *scripts* y de los externos con escala logarítmica en el eje Y. Como el número de *scripts* embebidos es muy superior al de externos, su distribución es muy similar a la del total, por lo que no se muestra.

Existen algunos sitios que utilizan más de 300 *scripts*, pero no son muchos, por lo que gran parte de la cola de la distribución es despreciable. También se han detectado un buen número de páginas que invocan repetidamente al mismo fichero de *script*. El uso de los mismos ficheros de *script* en varios dominios también ha resultado una buena forma de encontrar dominios con el mismo contenido.

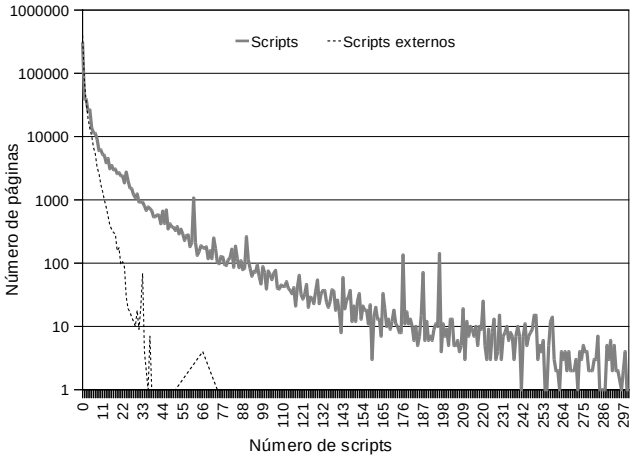

Fig. 2. Uso de ECMAScript en la Web española

Por otro lado, se ha realizado un estudio orientado a recopilar las librerías de *script* más populares. Así, se ha concluido que existe un grupo de 63 ficheros cuyos nombres aparecen más de 1.000 veces y que todos ellos siguen el estándar ECMAScript. La Tabla 4 muestra los primeros *scripts* de dicho grupo, con su número de ocurrencias y el número de dominios en los que se han detectado:

Nombre del fichero	Invocaciones	Dominios
show_ads.js	36.248	18.443
urchin.js	33.898	32.873
AC_RunActiveContent.js	29.746	28.526
swfobject.js	19.918	18.887
prototype.js	9.029	8.837
mootools.js	8.483	8.032
jquery.js	8.207	7.851
caption.js	5.947	5.916
scriptaculous.js	5.172	5.028
funciones.js	4.578	4.419

Tabla 4. Ficheros de script que se invocaron más de 1000 veces.

También se ha tratado de averiguar la función de las librerías más populares y de contar el número de dominios que hacían uso de las mismas. La Tabla 5 muestra el número de dominios y de invocaciones que hacen uso de ellas, agrupándolas según las funcionalidades para las que fueron diseñados los *scripts*.

Funcionalidad	Dominios	Invocaciones
Gestión de Flash y contenido activo	51.895	59.713
Recuento de visitas y generación de estadísticas	39.354	41.777
Dinamización del contenido con AJAX	28.819	41.767
Renderización del contenido/tratamiento de imágenes	22.185	24.598
Generación de menús	4.714	5.198
Tratamiento y validación de datos	4.376	6.586

Tabla 5. Dominios que emplean scripts con funcionalidades comunes.

Se ha llegado a la conclusión de que aunque hay muchas librerías en la red, casi todas se pueden clasificar en un número reducido de funcionalidades (generación de estadísticas, dinamización con AJAX, gestión de Flash, tratamiento de imágenes, etc.).

El uso de otros lenguajes de *script* es testimonial. Por ejemplo, tres de los cuatro *scritps* marcados como lenguaje TCL eran en realidad JavaScript marcado erróneamente. Por su parte, solo se han encontrado 1.769 llamadas a código VBScript, de las cuales únicamente 268 refieren a ficheros externos. De este último grupo, 246 contienen código relacionado con Flash (descarga del complemento, etc.).

Por último, se han estudiado informalmente los ficheros cuyo lenguaje no se ha podido detectar automáticamente. El resultado es que la mayoría de ellos son JavaScript, aunque puede ser generado dinámicamente (e.g.: aplicaciones CGI cuyo valor de retorno es el *script* que se ejecutará).

B. Formularios

Los formularios proporcionan el punto de entrada a la Web Oculta del lado servidor, por lo que también ha sido necesario estudiarlos en detalle. Se han encontrado 188.712 formularios en 124.865 dominios (21,6%). De ellos, 122.417 (un 64,9%) hacen su petición por POST y 48.443 (un 25,7%) hacen su petición por GET. Del resto, 17.779 (un 14,2%) asumen el valor por defecto (GET). Finalmente, 73 formularios establecen un valor no válido.

También se ha estudiado la distribución del número de formularios por dominio, obteniendo los datos mostrados en la Fig. 3:

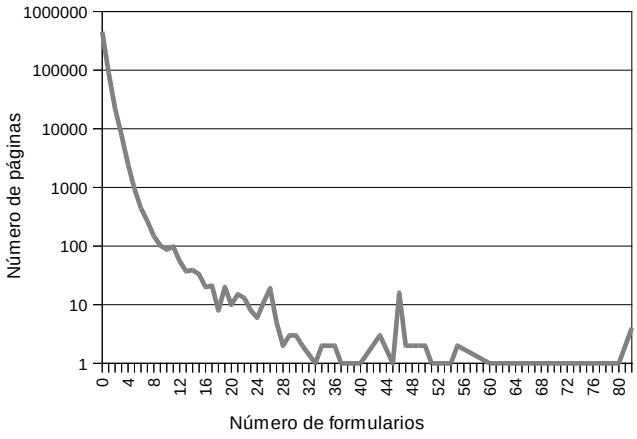

Fig. 3. Uso formularios en la Web española

La Tabla 6 muestra el uso de campos *password* en formularios. El porcentaje es relativo al número de dominios con formularios. El uso de este tipo de campos se asocia a funciones de autenticación, registro o cambio de contraseña.

Campos *password*	Formularios	Dominios	%
1 (autenticación)	26.918	25.832	20,7%
2 (registro)	251	239	0,2%
3 (cambio de clave)	33	33	<0,1%

Tabla 6. Formularios con campos de contraseña.

Se ha encontrado que 5.346 formularios en 4.861 dominios (3.9%) contienen un campo de texto y un botón. Se asume que su función es la de realizar búsquedas sencillas. En cuanto a las búsquedas e introducciones de datos más complejas, se ha hallado que 34.006 formularios en 31.288 dominios (25%) estaban conformados únicamente por

campos de texto y botones y que 126.095 formularios en 91.235 dominios (73%) contenían al menos dos elementos de las siguientes categorías: <input>, <select> y <textarea>.

En cuanto al uso de *scripts* en formularios, 743 dominios contienen formularios que incluyen la etiqueta "javascript:" en el atributo "action" y 735 dominios contienen formularios que incluyen los caracteres "()" en el "action", lo cual puede ser característico de una llamada a una función en un *script*.

También se han detectado 35.554 componentes típicos de formularios fuera de los mismos en 11.033 dominios (1,9%).

Por otro lado, existen 5.082 formularios en 3.742 dominios (0,6%) que no contienen componentes incluidos en la especificación de HTML 4.

C. Otras tecnologías

Las etiquetas <meta> pueden contener información de interés para los *crawlers* (exclusión de robots, redirecciones, *cookies*, etc.). La Tabla 7 muestra algunas funcionalidades para las que se han usado:

Función	**Dominios**	**%**
Refresco o redirección	22.064	3,8%
Refresco (sin URL para redirección)	1.346	0,2%
Estándar de exclusión de robots [16]	128.288	22,2%
Indicar las palabras clave	246.752	42,8%
Envío de *cookies*	26	<0,1%

Tabla 7. Uso de etiquetas <meta>.

Sin embargo, muchos buscadores no tienen en cuenta las palabras clave porque, tal y como se explica en [17], se pueden emplear para hacer *boosting*, es decir, para aumentar el *ranking* de la página y del dominio de forma injusta.

De los dominios que usan redirección, en 15.416 de los casos (2,6% del total de dominios), la página no incluía ningún enlace. En 15.039 casos (también un 2,6%) la página no incluía ni enlaces ni etiquetas <object>, lo cual descarta sitios 100% Flash.

Las aplicaciones Flash son otra dificultad a las que se han de enfrentar los *crawlers*. Se han encontrado etiquetas <object> en 89.911 dominios, lo que representa un 15,6% del total de la Web española. De ellos, 25.060 (un 4,3%) no presentaban ningún enlace mediante etiquetas <a>. En la mayor parte de esos casos, se trata de sitios web 100% Flash.

Por último, Se ha detectado que 257.084 dominios (un 44,6%) contienen elementos <link>. Estos elementos se suelen usar para hacer referencia a hojas de estilos, aunque también podrían referir a otro tipo de recursos (e.g.: *scripts*).

V. Conclusiones y trabajo futuro

Este artículo muestra los principales resultados de un análisis realizado sobre los sitios web de los dominios ".es" a fecha de 2009. En particular un 15,6% de los dominios presentan etiquetas <object> en la primera página, un 21,6% presentan formularios y un 46,2% contienen *scripts*.

La gran mayoría de las páginas emplean JavaScript como lenguaje de *scripting* y la mayoría de librerías de *script* responden a un conjunto reducido de propósitos: estadísticas, dinamización con AJAX, gestión de Flash, creación de menús, renderización de contenido, tratamiento de imágenes y validación de datos.

Por tanto, se puede concluir que una gran parte de los sitios web de los dominios ".es" hacen uso de tecnologías denominadas de Web Oculta, principalmente lenguajes de *scripting*. Por este motivo, están justificados los esfuerzos encaminados en dotar a los sistemas de *crawling* de mecanismos capaces de tratar con estas tecnologías para llegar al mayor número de documentos. El esfuerzo por interpretar dichos lenguajes debe estar dirigido a ECMAScript y en particular a JavaScript. Interpretar otros lenguajes como TCL o VBScript puede requerir un esfuerzo demasiado grande para un resultado poco significativo.

Como trabajo futuro se propone la realización de nuevos *crawlings* de los dominios ".es" para completar el estudio con una evolución de la Web española en los términos tratados en este artículo, para poder determinar de forma más precisa las tecnologías que deben de ser tratadas por los *crawlers*. Este estudio también puede ser interesante para analizar la frecuencia de refresco con la que los *crawlers* deberían recorrer ciertos sitios.

Agradecimientos

Este trabajo de investigación ha sido financiado por el Ministerio de Educación y Ciencia de España y los fondos FEDER de la Unión Europea (Proyecto TIN2009-14203).

El listado de dominios ".es" a partir del cual se ha realizado el *crawling* ha sido proporcionado por la Entidad Pública Empresarial Red.es.

Referencias

[1] M. Bergman. "The Deep Web. Surfacing Hidden Value," *Technical report, BrightPlanet LLC*. December 2000.
[2] M. Álvarez, F. Cacheda and A. Pan. "Análisis Macroscópico de los Dominios .es," *VIII Jornadas de Ingeniería Telemática (JITEL)*. 2009.
[3] K. C.-C. Chang, B. He, M. Patel, C. Li, and Z. Zhang. "Structured Databases on the Web: Observations and Implications," *SIGMOD Record*, vol. 33, no. 3, 2004.
[4] The size of the World Wide Web. http://www.worldwideWebsize.com/
[5] Netcraft. "March 2011 Web Servers Survey": http://news.netcraft.com/archives/category/web-server-survey/
[6] BuiltWith Technology Usage Statistics: http://trends.builtwith.com/
[7] Google - Web Authoring Statistics: http://code.google.com/intl/es-MX/webstats/index.html
[8] Internet Systems Consortium. "The ISC Domain Survey": http://www.isc.org/solutions/survey, 2011.
[9] Entidad Pública Empresarial Red.es: http://www.red.es
[10] VeriSign. "IPS Statistics - Internet-Profiling-Service": http://www.nic.at/en/uebernic/statistics/ips_statistics_informations/
[11] Standard ECMA-262: ECMAScript Language Specification: http://www.ecma-international.org/publications/standards/Ecma-262.htm
[12] M. Weideman and F. Schwenke. "The influence that JavaScript™ has on the visibility of a Website to search engines - a pilot study," *Information Research*, vol. 11, no. 4, July 2006.
[13] B. Wu and B.D. Davison. "Cloaking and Redirection: A Preliminary Study," *Proceedings of First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '05)*. 2005.
[14] CyberNeko HTML: http://sourceforge.net/projects/nekohtml/
[15] Scripting Media Types: http://www.rfc-editor.org/rfc/rfc4329.txt
[16] M. Koster "A Standard for Robot Exclusion," *Published online*. http://www.robotstxt.org/wc/norobots.htm, 1994.
[17] Z. Gyöngyi and H. Garcia-Molina. "Web Spam Taxonomy," *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '05)*. 2005.